Electronic Colloquium on Computational Complexity, Report No. 161 (2021)

On Worst-Case Learning in Relativized Heuristica

Shuichi Hirahara* Mikito Nanashima†

Abstract

A PAC learning model involves two worst-case requirements: a learner must learn *all* functions in a class on *all* example distributions. However, basing the hardness of learning on NP-hardness has remained a key challenge for decades. In fact, recent progress in computational complexity suggests the possibility that a weaker assumption might be sufficient for worst-case learning than the feasibility of worst-case algorithms for NP problems.

In this study, we investigate whether these *worst-case* requirements for learning are satisfied on the basis of only *average-case* assumptions in order to understand the nature of learning. First, we construct a strong worst-case learner based on the assumption that $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, i.e., in Heuristica. Our learner agnostically learns all polynomial-size circuits on all *unknown* $\mathsf{P/poly}$-samplable distributions in polynomial time, where the complexity of learning depends on the complexity of sampling examples. Second, we study the limitation of relativizing constructions of learners based on average-case heuristic algorithms. Specifically, we construct a powerful oracle such that $\mathsf{DistPH} \subseteq \mathsf{AvgP}$, i.e., every problem in PH is easy on average, whereas $\mathsf{UP} \cap \mathsf{coUP}$ and PAC learning on almost-uniform distributions are hard even for $2^{n/\omega(\log n)}$-time algorithms in the relativized world, which improves the oracle separation presented by Impagliazzo (CCC 2011). The core concept of our improvements is the consideration of a switching lemma on a large alphabet, which may be of independent interest. The lower bound on the time complexity is nearly optimal because Hirahara (STOC 2021) showed that $\mathsf{DistPH} \subseteq \mathsf{AvgP}$ implies that PH can be solved in time $2^{O(n/\log n)}$ under any relativized world.

*National Institute of Informatics, Japan. s_hirahara@nii.ac.jp

†Tokyo Institute of Technology, Japan. nanashima.m.aa@is.c.titech.ac.jp

ISSN 1433-8092

1 Introduction

Investigating the relationship between two fundamental tasks, namely computing and learning, has been a key research objective since Valiant introduced the probably approximately correct (PAC) learning model as a pioneering development in computational learning theory [Val84]. In the PAC learning model, a learner must learn *all* unknown target functions f computable by polynomial-size circuits on *all* unknown example distributions D in the following sense: the learner is given an accuracy parameter $\epsilon \in (0, 1]$ and passively collected examples of the form $(x, f(x))$, where each x is selected independently and identically according to D; then, the learner is asked to generate a good hypothesis h which is ϵ-close to f (i.e., $\Pr_{x\sim D}[h(x) \neq f(x)] \leq \epsilon$) with high probability.

A PAC learner must satisfy two worst-case requirements: it must be *distribution-free* and must learn *every* target function. The worst-case nature of PAC learning becomes more apparent in the equivalent model of Occam learning [BEHW87, Sch90]. In Occam learning, a learner is given an arbitrary set of examples and is asked to find a small hypothesis consistent with all the given examples. Clearly, this task can be formulated as a (worst-case) search problem in NP. The fundamental results of Blumer, Ehrenfeucht, Haussler, and Warmuth [BEHW87] and Schapire [Sch90] show that PAC learning and Occam learning are in fact equivalent.

Despite the worst-case nature of PAC learning and Occam learning, basing the hardness of these learning tasks on the hardness of NP has been a key challenges in computational learning theory for decades. The difficulty of proving the NP-hardness of learning has been explained in the work of Applebaum, Barak, and Xiao [ABX08], who showed that the NP-hardness of learning cannot be proved via a many-one reduction unless the polynomial hierarchy collapses. Given the lack of success in proving the NP-hardness of learning,[1] it is natural to ask whether PAC learning is "NP-intermediate." In this study, we investigate whether PAC learning is feasible in Heuristica [Imp95], i.e., a world in which NP is easy on average but hard in the worst case.

Recent progress in computational complexity has provided new insights into the relationship between learning and average-case complexity of NP. Several studies [CIKK16, CIKK17, HS17, ILO20] on natural proofs and the Minimum Circuit Size Problem (MCSP) have revealed that learning with respect to the uniform distribution can be formulated as an average-case NP problem. Carmosino, Impagliazzo, Kabanets, and Kolokolova [CIKK16] presented a generic reduction from the task of PAC learning with respect to the uniform distribution to a natural property [RR97]; a natural property is essentially equivalent to solving MCSP on average [HS17]. Therefore, these results imply that PAC learning for P/poly with respect to the uniform distribution is feasible under the assumption that MCSP $\in$ NP is easy on average.[2] Moreover, by combining this learning algorithm with inverters for distributional one-way functions [IL89], it can be shown that PAC learning with respect to every fixed samplable distribution on examples is feasible in Heuristica. Nanashima [Nan21] observed that, in Heuristica, it is possible to learn a target function chosen from a fixed samplable distribution with respect to every unknown example distribution. These results indicate that if either of the worst-case requirements on target functions or example distributions is weakened, then a polynomial-time learner for P/poly can be constructed from average-case (errorless) heuristics for NP problems.

[1] The main focus of this study is PAC learning for P/poly, i.e., the class of polynomial-size circuits. We mention that there are several NP-hardness results for proper PAC learning of restricted circuit classes [PV88].

[2] The learning algorithm of [CIKK16] requires a membership query. Ilango, Loff, and Oliveira [ILO20] showed that PAC learning with respect to the uniform distribution (without a membership query) is reduced to an average-case problem in NP.

1.1 Our Results

In this work, we present a PAC learner that satisfies the two worst-case requirements simultaneously in Heuristica. Under the assumption that NP admits average-case polynomial-time algorithms, we construct a polynomial-time learner that learns all polynomial-size circuits (P/poly) with respect to all *unknown* efficiently samplable example distributions. In fact, our learning algorithm learns polynomial-size circuits *agnostically*, i.e., even if a target function is not in P/poly, our learner outputs a hypothesis that is as good as the best hypothesis in P/poly.

Theorem 1 (informal)**.** *If* $\mathsf{DistNP} \subseteq \mathsf{AvgP}$ *(i.e.,* NP *is easy on average), then* P/poly *is agnostic learnable on all unknown* P/poly*-samplable distributions in polynomial time.*

Let us remark on several points. First, our learner works without knowing example distributions; however, it needs to know an upper bound on the complexity of example distributions. Second, the running time of our learner depends on the complexity of the concept class and example distributions. Note that the complexity of a learner does not depend on an example distribution in the standard PAC learning model. This is the only difference between the standard learning model and our learning model of Theorem 1; see Definition 2 for a precise definition. Third, most importantly in this work, the above-mentioned result is obtained by relativizing techniques, i.e., the above-mentioned theorem holds in the presence of any oracle.

Next, we consider the question of whether the standard PAC learner can be constructed in Heuristica. In other words, can we remove the condition of Theorem 1 that example distributions must be P/poly-samplable? We present strong negative answers by constructing "relativized Heuristica" in which there is no PAC learner with respect to almost-uniform distributions.

Theorem 2. *For any arbitrary small constant* $\epsilon > 0$*, there exists an oracle* $\mathcal{O}_\epsilon$ *such that*

(1) $\mathsf{DistPH}^{\mathcal{O}_\epsilon} \subseteq \mathsf{AvgP}^{\mathcal{O}_\epsilon}$ *and*

(2) $\mathsf{SIZE}^{\mathcal{O}_\epsilon}[n]$ *is not weakly learnable with membership queries in time* $O(2^{n/\omega(\log n)})$ *on all uniform distributions over* $S \subseteq \{0,1\}^n$ *such that* $|S| > 2^{(1-\epsilon)n}$.

This theorem shows that, unless we use some non-relativizing techniques, we cannot improve Theorem 1 for learning on almost-uniform example distributions even under the strong average-case assumption that $\mathsf{DistPH} \subseteq \mathsf{AvgP}$. Moreover, the hardness of learning holds even with the drastically weakened requirements: (a) weak learning (b) in sub-exponential time (c) with additional access to a membership query oracle.

In addition, we construct an oracle that separates the average-case complexity of PH from the worst-case complexity of $\mathsf{UP} \cap \mathsf{coUP}$ with the best possible parameters on time complexity.

Theorem 3. *There exists an oracle* $\mathcal{O}$ *such that*

$$\textit{(1)}\ \mathsf{DistPH}^{\mathcal{O}} \subseteq \mathsf{AvgP}^{\mathcal{O}} \ \textit{and (2)}\ \mathsf{UP}^{\mathcal{O}} \cap \mathsf{coUP}^{\mathcal{O}} \not\subseteq \mathsf{BPTIME}^{\mathcal{O}}[2^{n/\omega(\log n)}].$$

Furthermore, for all $k \in \mathbb{N}$ *and constants* $a > 0$*, there exists an oracle* $\mathcal{O}_{k,a}$ *such that*

$$\textit{(1)}\ \mathsf{Dist}\Sigma_k^{p\mathcal{O}_{k,a}} \subseteq \mathsf{AvgP}^{\mathcal{O}_{k,a}} \ \textit{and (2)}\ \mathsf{UP}^{\mathcal{O}_{k,a}} \cap \mathsf{coUP}^{\mathcal{O}_{k,a}} \not\subseteq \mathsf{BPTIME}^{\mathcal{O}_{k,a}}[2^{an/\log n}].$$

This result significantly improves the previous oracle construction of Impagliazzo [Imp11], who proved that there exist a constant $\alpha > 0$ and an oracle $\mathcal{O}$ such that

$$(1)\ \mathsf{DistNP}^{\mathcal{O}} \subseteq \mathsf{AvgP}^{\mathcal{O}} \text{ and } (2)\ \mathsf{UP}^{\mathcal{O}} \cap \mathsf{coUP}^{\mathcal{O}} \not\subseteq \mathsf{BPTIME}^{\mathcal{O}}[2^{n^\alpha}].$$

In Theorem 3, we improve this oracle construction in the following two aspects. First, the worst-case lower bound is improved from $O(2^{n^\alpha})$ to $2^{n/\omega(\log n)}$. Second, the feasibility of the average-case computation is improved from DistNP to DistPH. The core concept of our improvements is to consider a switching lemma on a large alphabet, which may be of independent interest.

Recently, Hirahara [Hir21] presented the first nontrivial worst-case-to-average-case connection for PH:

Theorem 4 ([Hir21]). • *If* $\mathsf{DistPH} \subseteq \mathsf{AvgP}$, *then* $\mathsf{PH} \subseteq \mathsf{DTIME}[2^{O(n/\log n)}]$; *and*

• *If* $\mathsf{Dist}\Sigma^p_{k+1} \subseteq \mathsf{AvgP}$, *then* $\Sigma^p_k \subseteq \mathsf{DTIME}[2^{O(n/\log n)}]$ *for each* $k \in \mathbb{N}$.

This result is proved by a relativizing proof technique (see Appendix A for the details). Therefore, the time complexity $2^{n/\omega(\log n)}$ given in Theorem 3 is nearly optimal for PH and completely optimal for Σ^p_k.

1.2 Related Work

Impagliazzo and Levin [IL90] constructed another type of learner called universal extrapolation under the assumption that there is no one-way function (i.e., in Pessiland), where the learner approximates the appearance probability of a given string generated according to some unknown distribution samplable by an efficient *uniform* algorithm. The agnostic learner of Theorem 1 can learn a polynomial-size circuit with respect to unknown distributions samplable by efficient *non-uniform* algorithms; however, it requires a stronger assumption that NP is easy on average (i.e., in Heuristica). Li and Vitányi [LV91] implicitly developed a PAC learner on simple distributions that contain $\mathsf{P/poly}$-computable distributions in Heuristica. However, their learner requires an additional example oracle on a specific distribution (i.e., the time-bounded universal distribution). In another line of research, several cryptographic primitives have been constructed on the basis of the hardness of learning linear functions on the uniform distribution in noisy settings (e.g., [Reg09, DP12]). Theorem 1 can be regarded as a step toward constructing such cryptographic primitives based on the weaker hardness assumption of learning in general settings. Several studies [Dan16, DSS16, Vad17] have shown the hardness of learning for various central concept classes (e.g., polynomial-size DNFs) under the average-case hardness of constraint satisfaction problems, in contrast to our work.

Regarding Theorems 3 and 2, Xiao [Xia09] constructed an oracle that separates the hardness of learning on the uniform distribution from the non-existence of an auxiliary-input one-way function. Since the hardness of learning on the uniform distribution implies that $\mathsf{DistNP} \not\subseteq \mathsf{AvgP}$, the scope is different from ours. Watson [Wat12] constructed an oracle that rules out black-box reductions from worst-case UP to average-case PH. Our relativization barrier is more general in that the proof of Theorem 4 uses non-black-box reductions and is not captured by the oracle construction of [Wat12]. Another line of research [FF93, BT06b, AGGM06, GV08, ABX08, HMX10, BL13, BB15, LV16, HW20] rules out restricted types of black-box reductions from NP-hard languages to several average-case notions under the assumption that PH does not collapse, which is not comparable with our relativization result.

2 Overview of Proof Techniques

In this section, we present an overview of our proof techniques.

2.1 Agnostic Learner in Heuristica

Here, we explain the ideas for constructing the agnostic learner of Theorem 1 under the assumption that $\mathsf{DistNP} \subseteq \mathsf{AvgP}$. Our proofs are based on two lemmas, which we explain below.

The first lemma is the worst-case to average-case connection developed in [Hir18, Hir20] for the problem of computing the time-bounded Kolmogorov complexity. Fix a prefix-free universal Turing machine U_0 arbitrarily. For each $t \in \mathbb{N}$ and $x \in \{0,1\}^*$, the *t-time-bounded Kolmogorov complexity* of x is defined as

$$K^t(x) = \min_{p \in \{0,1\}^*} \{|p| : U_0(p) \text{ outputs } x \text{ in } t \text{ steps}\}.$$

We also define $K(x)$ by $K(x) = \lim_{t\to\infty} K^t(x)$. Intuitively, $K^t(x)$ represents the minimum length of a program that outputs x in time t. It was shown in [Hir20] that the time-bounded Kolmogorov complexity can be efficiently approximated in the worst case under the assumption that NP is easy on average.

Lemma 1 ([Hir20])**.** *If $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, then there exist a polynomial τ and an algorithm $\mathsf{ApproxK}_\tau$ that is given $(x, 1^t)$, where $x \in \{0,1\}^*, t \in \mathbb{N}$, and outputs an integer $s \in \mathbb{N}$ in polynomial time to satisfy*

$$K^{\tau(|x|+t)}(x) - \log \tau(|x|+t) \leq s \leq K^t(x).$$

The second lemma is a recent characterization of learnability based on a task called random-right-hand-side refutation (RRHS-refutation), first introduced by Vadhan [Vad17]. RRHS-refutation enables us to characterize the feasibility of PAC learning. This characterization was later extended by Kothari and Livni [KL18] to a characterization of agnostic learning, which we call *correlative RRHS-refutation*.

We briefly explain the task of correlative RRHS-refutation[3]: For a randomized function $f\colon \{0,1\}^n \to \{0,1\}$, a concept class $\mathscr{C}$, and a distribution D on $\{0,1\}^n$, we define a correlation $\mathrm{Cor}_D(f,\mathscr{C}) \in [-1,1]$ between f and $\mathscr{C}$ with respect to D by

$$\mathrm{Cor}_D(f,\mathscr{C}) := \max_{c\in\mathscr{C}_n} \mathrm{E}_{f,x\leftarrow D}\left[(-1)^{f(x)} \cdot (-1)^{c(x)}\right] = 2 \cdot \max_{c\in\mathscr{C}_n} \Pr_{f,x\leftarrow D}[f(x) = c(x)] - 1.$$

Roughly speaking, correlative RRHS-refutation for $\mathscr{C}$ on a class $\mathcal{D}$ of example distributions is a task of distinguishing the following two cases with high probability: on input $\epsilon \in (0,1]$, (i) a "correlative" case where samples are chosen identically and independently according to $EX(f, D_n)$ for $D_n \in \mathcal{D}_n$ and a randomized function f such that $\mathrm{Cor}_{D_n}(f,\mathscr{C}) \geq \epsilon$; and (ii) a "random" case where samples are chosen identically and independently according to $EX(f_R, D_n)$ for $D_n \in \mathcal{D}_n$ and a truly random function f_R. Kothari and Livni [KL18] showed that for every concept class $\mathscr{C}$, $\mathscr{C}$ is correlatively RRHS-refutable in polynomial time iff $\mathscr{C}$ is agnostic learnable in polynomial time. In light of this characterization, our goal is to perform correlative RRHS-refutation using an approximation algorithm for time-bounded Kolmogorov complexity.

Correlative RRHS-refutation on Shallow Sampling-Depth Distributions

Now, we present a proof idea for constructing a correlative RRHS-refutation algorithm using an approximation algorithm $\mathsf{ApproxK}_\tau$ for time-bounded Kolmogorov complexity. Our refutation algorithm operates as follows:

[3]In this paper, we use a different term (i.e., correlative RRHS-refutation) to refer to "refutation" in the original paper [KL18] in order to distinguish it from "RRHS-refutation" in [Vad17] and other refuting tasks for random CSPs.

1. For a given sample $S = ((x^{(i)}, b^{(i)}))_{i=1}^{m}$, let $X = x^{(1)} \circ \cdots \circ x^{(m)}$ and $b = b^{(1)} \circ \cdots \circ b^{(m)}$, where $\circ$ denotes the concatenation of strings.
2. Use $\mathsf{ApproxK}_\tau$ to approximate $K^t(X)$ and $K^{t'}(X \circ b)$ for some time bounds t and t', respectively. Let s and s' denote the respective approximated values.
3. If $\Delta = s' - s$ is less than some threshold T, then output "correlative"; otherwise, output "random".

We explain why this algorithm distinguishes the "correlative" case and the "random" case. In the former case, samples X and b are generated by a target function f such that $\mathrm{Cor}_D(f, \mathscr{C}) \geq \epsilon$. Thus, the best concept $c^* \in \mathscr{C}$ satisfies that $\Pr_{f,x \leftarrow D}[c^*(x) \neq f(x)] \leq 1/2 - \epsilon/2$. Using this fact, we claim that the t'-time-bounded Kolmogorov complexity of $X \circ b$ is small for a properly large t'. Let $e \in \{0,1\}^m$ denote the string that indicates the difference between c^* and f, i.e., the i-th bit of e is $b^{(i)} \oplus c^*(x^{(i)})$ for every $i \in [m]$. Using the best concept c^*, a program d_X that describes X, and the string e that indicates an "error", we can describe the string $X \circ b$ by the following procedure: (1) compute X, (2) compute $b^* = c^*(x^{(1)}) \cdots c^*(x^{(m)})$ by applying c^* to each input $x^{(i)}$ contained in X, and (3) compute b (and output $X \circ b$) by taking bit-wise XOR between b^* and e. The length of the description of this procedure is bounded above by

$$|d_X| + |c^*| + |(\text{a description of } e)| + O(1) \leq s + \ell_{\mathscr{C}}(n) + (1 - \Omega(\epsilon^2)) \cdot m,$$

with high probability, where $\ell_{\mathscr{C}}(n)$ is the length of the representation of the n-input functions in $\mathscr{C}$. Therefore, $\Delta = s' - s$ is at most $\ell_{\mathscr{C}}(n) + (1 - \Omega(\epsilon^2)) \cdot m$ in a "correlative" case.

Thus, if $\Delta \approx m$ holds with high probability in a "random" case, then the algorithm distinguishes a "random" case from a "correlative" case by taking sufficiently large m with respect to $n, \ell_{\mathscr{C}}(n)$, and ϵ^{-1}. It seems reasonable to expect that $\Delta \approx m$ because b is a truly random string of m bits selected independently of X. However, in general, this might not hold for the following two technical reasons. First, we need nearly m bits to describe b with high probability; however, such b might help generate X in a time-bounded setting. Second, we must choose a time bound t' larger than t to ensure the upper bound on Δ in a "correlative" case, and this might also reduce the cost of generating X.

To analyze the case in which Δ becomes large, we consider the expected value of the *computational depth* of samples. Antunes, Fortnow, van Melkebeek, and Vinodchandran [AFvV06] introduced the notion of the t-time-bounded computational depth of $x \in \{0,1\}^*$ (where $t \in \mathbb{N}$), which is defined as $K^t(x) - K(x)$. Hirahara [Hir21] extended this notion to the (t,t')-time-bounded computational depth of $x \in \{0,1\}^*$ (where $t, t' \in \mathbb{N}$ with $t' > t$), which is defined as $K^t(x) - K^{t'}(x)$. Here, we further generalize these notions as follows:

Definition 1 (Sampling-depth functions)**.** *Let $t, t' \in \mathbb{N}$ such that $t' > t$. For a class $\mathcal{D}$ of example distributions, we define a (t,t')-sampling-depth function $sd_{\mathcal{D}}^{t,t'} = \{sd_{\mathcal{D},n}^{t,t'}\}_{n \in \mathbb{N}}$ where $sd_{\mathcal{D},n}^{t,t'} \colon \mathbb{N} \to \mathbb{R}_{\geq 0}$ by*

$$sd_{\mathcal{D},n}^{t,t'}(m) = \max_{D \in \mathcal{D}} \mathrm{E}_{X_D}\left[K^t(X_D) - K^{t'}(X_D)\right],$$

where $X_D = x^{(1)} \circ \cdots \circ x^{(m)}$, and each $x^{(i)}$ is selected identically and independently according to D.

We verify that if the sampling depth of example distributions is small, then Δ is large. We remark that Δ could become small because (i) the random string b and (ii) the larger time bound $t' (> t)$ could help generate X. However, if the sampling depth of the example distribution is small for t and t', the second case does not occur because $K^{t'}(X)$ is close to $K^t(X)$ with high probability. To show that the first case does not occur, we apply the weak symmetry of information, proved by

Hirahara [Hir21] under the assumption that NP is easy on average. Informally speaking, the weak symmetry of information states that for any time bound $t \in \mathbb{N}$ and string X, and for a random string b, $K^t(X \circ b)$ is larger than $K^{t'}(X) + |b|$ for some large $t' > t$ with high probability over the choice of b. By the weak symmetry of information and the small sampling depth of the example distribution, we can show that $K^{t'}(X \circ b)$ is large compared to $K^t(X) + |b|$, i.e., the additional random string b does not help generate X so much.

To show Theorem 1, we will also observe that the sampling-depth function of a P/poly-samplable distribution is logarithmically small. Roughly speaking, this follows from the fact that samples selected according to a P/poly-samplable distribution have nearly optimal encoding with an efficient decoder, which can be proved using the techniques developed in [AF09, AGvM+18, Hir21]. In other words, the term $\mathrm{E}[K^t(X_D)]$ in the definition of a sampling-depth function is nearly close to $mH(D) \approx \mathrm{E}[K(X_D)]$ for a sufficiently large t, where $H(D)$ is the entropy of D.

2.2 Oracle Separation

In this subsection, we present our proof ideas for Theorems 2 and 3.

Before presenting our key idea to show Theorem 3, we first explain the idea applied in [Imp11] and the reason why it is not sufficient for the improved lower bound $2^{\Omega(n/\log n)}$.

The oracle $\mathcal{O}$ constructed in [Imp11] consists of the following two oracles: a random permutation $\mathcal{F} = \{\mathcal{F}_n\}_{n\in\mathbb{N}}$, where $\mathcal{F}_n \colon \{0,1\}^n \to \{0,1\}^n$, and a restricted NP-oracle $\mathcal{A}$. The oracle $\mathcal{A}$ takes a nondeterministic oracle machine $M^?$, $x \in \{0,1\}^*$, and 1^{T^4}, where $T \in \mathbb{N}$, as input and simulates $M^{\mathcal{F}+\mathcal{A}}(x)$ in T steps. We remark that the simulation overhead T^4 in $\mathcal{A}$ is crucial for preventing circular calls for $\mathcal{A}$. The purpose of $\mathcal{F}$ is to make UP∩coUP hard by considering its inverting problem, and the purpose of $\mathcal{A}$ is to make DistNP easy on average in the relativized world. A challenging task in the construction is to preserve the worst-case hardness of NP, even in the presence of the restricted NP-oracle $\mathcal{A}$.

To satisfy this requirement, the key idea applied in [Imp11] is to let $\mathcal{A}$ reveal the values of $\mathcal{F}$ gradually according to the time bound T. The execution of a nondeterministic machine $M^{\mathcal{O}}$ is represented as a disjunctive normal form (DNF) formula in variables $F_{x,y}$ for $x, y \in \{0,1\}^*$ with $|x| = |y|$ (referred to as matching variables), which expresses the connection specified by $\mathcal{F}$ (i.e., $F_{x,y} = 1$ iff $\mathcal{F}(x) = y$). Impagliazzo's idea is to apply random restrictions to these matching variables repeatedly on the choice of $\mathcal{F}$, i.e., to determine the values of $\mathcal{F}$ in multiple steps. In the execution of $\mathcal{A}$, we determine the disclosure levels for $\mathcal{F}$ as follows. On input $(M, x, 1^{T^4})$, $\mathcal{A}$ applies only the first $i := 2^{-1} \log\log T$ restrictions to the DNF formula ϕ_M corresponding to $M^?(x)$ (where i is selected so that circular calls for $\mathcal{A}$ will not occur). If ϕ_M becomes a constant by these restrictions, then $\mathcal{A}$ returns the same constant; otherwise, $\mathcal{A}$ returns "?". We remark that, whenever $\mathcal{A}(M, x, 1^{T^4})$ returns some constant, the answer by $\mathcal{A}$ is consistent with the answer of $M^{\mathcal{O}}(x)$ executed in T steps. To solve the $\mathsf{NP}^{\mathcal{O}}$ problem $L^{\mathcal{O}}$ determined by a polynomial-time nondeterministic machine $M^{\mathcal{O}}$ on average, we query $(M, x, 1^{T^4})$ to $\mathcal{A}$ for an input x and a sufficiently large T with respect to the time bound of M and return the answer from $\mathcal{A}$. When the instance x is selected by some efficient sampler $S^{\mathcal{O}}$, $S^{\mathcal{O}}$ cannot access $\mathcal{F}$ at high disclosure levels with high probability, and the instance x is independent of such values of $\mathcal{F}$. In this case, the average-case easiness for NP follows from *the switching lemma for DNFs on matching variables*. Roughly speaking, the lemma shows that the output of any small depth DNF formula on matching variables is fixed to a constant with high probability by applying a random restriction. Since the simulation of M by $\mathcal{A}$ is regarded as the application of a random restriction to ϕ_M, the switching lemma guarantees that the value of ϕ_M is determined with high probability, and it must be the correct answer for $L^{\mathcal{O}}$. Meanwhile, inverting $\mathcal{F}$ remains hard in the worst case as long as the inverting algorithms do not have sufficient resources

to fully access $\mathcal{F}$.

The bottleneck in the above-mentioned construction lies in the bad parameters of the switching lemma for matching variables. Let N be the number of unassigned entries of $\mathcal{F}_n$ (for some $n \in \mathbb{N}$) at some stage when selecting random restrictions. To obtain the nontrivial bound on the failure probability of $\mathcal{A}$ (i.e., the probability that ϕ_M does not become a constant by a random restriction) by applying the switching lemma for matching variables, we need to additionally assign at least $N - \sqrt{N}$ entries of $\mathcal{F}_n$. To obtain the lower bound $t(n) = 2^{\Omega(n/\log n)}$ in the result, we need to apply such random restrictions $i_{\max}(n) := 2^{-1} \log \log t(n)$ times to prevent $t(n)$-time algorithms from accessing all random restrictions (i.e., full access to $\mathcal{F}$) by $\mathcal{A}$. In these settings of parameters, all the values of $\mathcal{F}$ are assigned before applying random restrictions $i_{\max}(n)$ times. In other words, $t(n)$-time algorithms can access to all the information about $\mathcal{F}$ by $\mathcal{A}$, which is sufficient to invert $\mathcal{F}$ efficiently.

Switching Lemma on General Domains

Now, we present the key idea for improving the lower bound. In this section, for simplicity, we focus on the case of separation from DistNP.

The key idea for the improvement is to apply a switching lemma *on general domains* instead of the switching lemma for matching variables, where the variables are separated into several blocks and take different alphabets in different blocks. The size of the alphabets and the probability of random restrictions also vary among the blocks. We first present the details of the switching lemma and then explain the oracle construction and the importance of large alphabets.

Let Σ be a finite set of alphabets. For a variable x that takes a value in Σ, we define a literal on x as a condition taking either of the following forms for some $a \in \Sigma$: (i) $x = a$ or (ii) $x \neq a$. Using these generalized literals, we define DNFs, conjunctive normal form formulas (CNFs), and circuits of general domains as the usual ones of a binary domain.

For $p \in [0, 1]$ and a set V of variables on Σ, we define a p-random restriction $\rho \colon V \to \Sigma \cup \{*\}$ by the following procedure. First, we select a random subset $S \subseteq V$ of size $\lfloor p|V| \rfloor$ uniformly at random. Then, we set $\rho(x) = *$ (which represents "unassigned") for $x \in S$ and assign a uniformly random value $\rho(x)$ from Σ for each $x \in V \setminus S$. For partial assignments ρ_1 to variables V_1 and ρ_2 to variables V_2, we use the notation $\rho_1\rho_2$ to represent the composite restriction to $V_1 \cup V_2$. Then, our technical lemma is stated as follows.

Lemma 2. *For $m \in \mathbb{N}$, let $\Sigma_1, \ldots, \Sigma_m$ be finite sets of alphabets, and let $V_1, \ldots, V_m$ be disjoint sets of variables, where each variable in V_i takes a value in Σ_i. For each $i \in [m]$, let ρ_i be a p_i-random restriction to V_i, where $p_i \in [0, 1]$. Then, for any t-DNF ϕ on the variables in $V_1 \cup \ldots \cup V_m$ and $k \in \mathbb{N}$, we have*

$$\Pr_{\rho_1,\ldots,\rho_m} [\phi|_{\rho_1 \ldots \rho_m} \textit{ is not expressed as } k\textit{-CNF }] \leq O\left(mt \cdot \max_{i \in [m]} p_i |\Sigma_i|^2\right)^k.$$

Tight Separation between UP ∩ coUP and DistNP

Here, we present the oracle construction for separating UP ∩ coUP and DistNP and explain why large alphabets on the switching lemma are important for the tight lower bound $t(n) = 2^{\Omega(n/\log n)}$.

In our construction, an oracle $\mathcal{O}$ consists of two oracles $\mathcal{V}$ and $\mathcal{A}$, where $\mathcal{V}$ makes UP ∩ coUP hard and $\mathcal{A}$ makes DistNP easy on average in the relativized world. Further, $\mathcal{V}$ and $\mathcal{A}$ are determined by the internal random function $f = \{f_n\}_{n \in \mathbb{N}}$, where $f_n \colon \{0, 1\}^n \to \Sigma_n$, and Σ_n is a subexponentially large alphabet in n. For each $n \in \mathbb{N}$, we select the random function f_n by repeatedly applying $p(n)$-random

restrictions $i_{\max}(n) := \Theta(\log\log t(n))$ times and determine the disclosure levels from 1 to $i_{\max}(n)$ (i.e., full access to f_n) on the execution of $\mathcal{A}$. We define $\mathcal{V}$ by $\mathcal{V}(x,y) = 1$ if $F(x) = y$; otherwise, $\mathcal{V}(x,y) = 0$. We also define the restricted NP oracle $\mathcal{A}$ similarly to the previous construction. The easiness of DistNP follows from the switching lemma on general domains (Lemma 2).

In our oracle construction, computing f_n is hard for $t(n)$-time algorithms because any $t(n)$-time algorithm cannot obtain any information of f of the highest disclosure level from $\mathcal{A}$ by the choice of $i_{\max}(n)$. In fact, we can obtain the lower bound close to $|\Sigma_n|$ on the time complexity of computing f_n, even with access to $\mathcal{V}$. The lower bound for $\mathsf{UP}^{\mathcal{O}} \cap \mathsf{coUP}^{\mathcal{O}}$ holds because computing f_n is reducible to the following language $L^{\mathcal{O}}$ in $\mathsf{UP}^{\mathcal{O}} \cap \mathsf{coUP}^{\mathcal{O}}$:

$$L^{\mathcal{O}} = \{(x,i) : n \in \mathbb{N}, x \in \{0,1\}^n, i \in [n], \text{ and } \exists y \in \Sigma_n \text{ s.t. } f_n(x) = y \text{ and } \langle y \rangle_i = 1\},$$

where $\langle y \rangle$ denotes a (proper and unique) binary representation of $y \in \Sigma_n$.

Next, we explain the importance of large alphabets. It is natural to attempt to use a standard switching lemma on binary alphabets because the parameters achievable by such a switching lemma are significantly better than those achievable by a switching lemma on matching variables. We explain below why this approach is insufficient to obtain the tight lower bound. Let $f_n \colon \{0,1\}^n \to \{0,1\}^{\mathsf{poly}(n)}$ be a random function constructed by repeatedly applying the standard $p(n)$-random restrictions on each bit of $f_n(x)$ for every $x \in \{0,1\}^n$. Note that the output length of f_n must be at most $\mathsf{poly}(n)$ in order to make sure that $L^{\mathcal{O}} \in \mathsf{UP}^{\mathcal{O}} \cap \mathsf{coUP}^{\mathcal{O}}$. There are two conflicting requirements on the probability $p(n)$.

On one hand, to obtain the lower bound $t(n) = 2^{\Omega(n/\log n)}$, we need to let $p(n)$ be subexponentially small for the following reasons. Consider the case in which an $\mathsf{NP}^{\mathcal{O}}$ problem $L^{\mathcal{O}}$ is determined by a polynomial-time nondeterministic machine $M^{\mathcal{O}}$, and query $M^{\mathcal{O}}$ to $\mathcal{A}$ for some time bound $T = \mathsf{poly}(n)$ to solve $L^{\mathcal{O}}$ on average. For each instance $x \in \{0,1\}^n$, $M^{\mathcal{O}}(x)$ may access $f_{n'}$ by $\mathcal{A}$ for some $n' \approx t^{-1}(T)$ such that $i_{\max}(n')$ $(= \Theta(\log\log t(n')))$ is slightly larger than the disclosure level $\Theta(\log\log T)$, which is accessible to $M^{?}$. In this case, random restrictions for $f_{n'}$ are applied in the simulation of $M^{\mathcal{O}}(x)$ by $\mathcal{A}$. To bound the failure probability of $\mathcal{A}$ above by $1/q(n)$ for some $q(n) = \mathsf{poly}(n)$ by the switching lemma, we need to select $p(n)$ to satisfy $p(n') \approx p(t^{-1}(T)) = p(t^{-1}(\mathsf{poly}(n))) \leq 1/q(n)$. To satisfy this requirement, we need to select a subexponentially small $p(n)$.

On the other hand, $p(n)$ must be at least $1/\mathsf{poly}(n)$ in order to show the worst-case lower bound on the time complexity of $L^{\mathcal{O}}$. In general, it is possible to obtain approximately 2^{d_n} as the lower bound on the time complexity of computing f_n, where d_n is the maximum number of $*$ contained in $f_n(x)$ for some $x \in \{0,1\}^n$ at the $(i_{\max}(n)-1)$ disclosure level. However, when we apply random restrictions for binary variables with a sub-polynomially small probability $p(n) \leq n^{-\omega(1)}$, it holds that $d_n = o(n/\log n)$ with high probability; hence, we cannot obtain the desired lower bound $2^{\Omega(n/\log n)}$ by just using a switching lemma on binary alphabets.

The switching lemma on general domains insists that the size of the alphabets only affects the probability of a random restriction multiplicatively. Thus, we can select subexponentially many alphabets even for subexponentially small $p(n)$ without affecting the failure probability. This yields sufficient $*$s in $f_n(x)$ at the $(i_{\max}(n)-1)$ disclosure level, and interestingly, it yields the tight subexponential lower bound for $\mathsf{UP} \cap \mathsf{coUP}$.

We remark that we can extend the above-mentioned argument to the case of the separation between $\mathsf{UP} \cap \mathsf{coUP}$ and DistPH by extending Lemma 2 to constant-depth circuits as the standard switching lemma.

Extending the Hardness from UP ∩ coUP to Learning

To extend the hardness result to learning, we change the oracle $\mathcal{V}$ in the above-mentioned construction to a new oracle $\mathcal{F}$ determined as follows. In addition to the random function $f = \{f_n\}_{n\in\mathbb{N}}$, where $f_n\colon \{0,1\}^n \to \Sigma_n$, we select an internal random function $g = \{g_n\}_{n\in\mathbb{N}}$, where $g_n\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ by repeatedly applying random restrictions. Intuitively, we use g as the target function $g_z(x) := g(z,x)$ for each $z \in \{0,1\}^*$ and f as the pair of locks and keys to access g through the oracle $\mathcal{F}$. Specifically, we define the oracle $\mathcal{F}$ by $\mathcal{F}(z,y,x) = g_z(x)$ if $f(z) = y$; otherwise, $\mathcal{F}(z,y,x) = 0$.

Then, we construct an oracle $\mathcal{O}$ consisting of $\mathcal{F}$ and the restricted NP oracle $\mathcal{A}$. The average-case easiness of DistNP follows from the switching lemma on general domains in a similar way, where we identify each entry $f_n(z)$ (for $z \in \{0,1\}^n$) with a variable on Σ_n and identify each entry $g(z,x)$ (for $z, x \in \{0,1\}^n$) with a binary variable.

Now, we present the proof sketch of the hardness of learning. We consider the following concept class $\mathscr{C}^{\mathcal{O}} = \{h_{z,y} : h_{z,y}(x) = \mathcal{F}(z,y,x) \text{ for } z \in \{0,1\}^n \text{ and } y \in \Sigma_n\}$. Since a worst-case learner L for $\mathscr{C}^{\mathcal{O}}$ learns $\mathcal{F}(z,y,x)$ for all $z \in \{0,1\}^n$ and $y \in \Sigma_n$, such an L must learn g_z for all $z \in \{0,1\}^n$. Note that the learner L can access $\mathcal{F}$ but not g_z through $\mathcal{F}$ unless the key $f(z)$ is identified.

We will show the upper bound on the probability that L succeeds in learning $\mathscr{C}^{\mathcal{O}}$ without sufficient resources for full access to f and g by $\mathcal{A}$. There are the following two cases for L: (1) L finds $f(z)$ for all z with notable probability, or (2) L learns g_z without identifying $f(z)$ for some z. In the former case, L essentially succeeds in computing f, which must be hard in the worst case, as discussed in the case of $\mathsf{UP} \cap \mathsf{coUP}$. In the latter case, if we consider the case of learning g_z on the uniform distribution over unrevealed entries of g_z at the $(i_{\max}(n)-1)$ disclosure level, then L cannot distinguish the value of g_z with a truly random value on the support of the example distribution even with access to $\mathcal{A}$. Thus, L cannot learn g_z even weakly on such an example distribution. In fact, we can show that there exists an index z with high probability such that the value $f(z)$ is unassigned and many $*$s remain in the truth table of g_z at the $(i_{\max}(n)-1)$ disclosure level. This yields the subexponential lower bound of weak learning on almost-uniform distributions.

2.3 Organization of this Paper

The remainder of this paper is organized as follows. In Section 3, we introduce the preliminaries for our formal arguments. In Section 4, we present our agnostic learner and analyze its capability. In Section 5, we present the switching lemma on general domains. By applying the switching lemma, we show the oracle separation between $\mathsf{UP} \cap \mathsf{coUP}$ and DistPH in Section 6 and that between worst-case learning and DistPH in Section 7.

3 Preliminaries

For each $n \in \mathbb{N}$, let $[n] = \{1, \ldots, n\}$. For a distribution D, we use the notation $x \leftarrow D$ to denote a random sampling x according to D. For a finite set S, we also use the notation $x \leftarrow_u S$ to denote the uniform sampling from S. For $x \in \{0,1\}^*$, let $D(x) \in [0,1]$ be the probability that x is generated according to D. For each distribution D and $m \in \mathbb{N}$, let D^m denote the distribution of $x_1 \circ \cdots \circ x_m$, where $x_1, \ldots, x_m \leftarrow D$. For any distribution D, let $H(D)$ denote the Shannon entropy of D.

For a randomized algorithm A using $r(n)$ random bits on an n-bit input, we use $A(x;s)$ to refer to the execution of $A(x)$ with a random tape s for $x \in \{0,1\}^n$ and $s \in \{0,1\}^{r(n)}$.

In this paper, we assume basic knowledge of probability theory, including the union bound, Markov's inequality, Hoeffding's inequality, and the Borel–Cantelli lemma. We also use the following concentration inequality to select a random subset. For correctness, we present the formal proof in Appendix B.

Lemma 3. *Let U be a universe of size N, and let $Z \subseteq U$ be an arbitrary subset of size M ($\leq N$). Let $S \subseteq U$ be a random subset of size n. Then, for any $\gamma \in (0,1)$, we have*

$$\Pr_T\left[\left||S \cap Z| - \frac{M}{N}n\right| > \gamma \cdot \frac{M}{N}n\right] < 2e^{-2\gamma^2 \cdot (\frac{M}{N})^2 \cdot n}.$$

3.1 Learning Models

A concept class is defined as a subset of Boolean-valued functions $\{f : \{0,1\}^n \to \{0,1\} : n \in \mathbb{N}\}$. Roughly speaking, the goal of learners is to learn all functions in a concept class from passively collected data.

For any concept class $\mathscr{C}$, we use the notation $\mathscr{C}_n$ to represent $\mathscr{C} \cap \{f : \{0,1\}^n \to \{0,1\}\}$ for each $n \in \mathbb{N}$. We assume that every concept class $\mathscr{C}$ has a binary encoding of the target functions and an evaluation $C : \{0,1\}^* \times \{0,1\}^* \to \{0,1\}$ satisfying $C(f,x) = f(x)$ for each $n \in \mathbb{N}$, $f \in \mathscr{C}_n$, and $x \in \{0,1\}^n$. In this paper, we consider only polynomially evaluatable concept classes, i.e., classes that have polynomial-size encodings and polynomial-time computable evaluation functions. For every $\mathscr{C}$, we use the notation $\ell_{\mathscr{C}}$ to refer to a polynomial $\ell_{\mathscr{C}} : \mathbb{N} \to \mathbb{N}$ such that each $f \in \mathscr{C}_n$ has a binary encoding of length at most $\ell_{\mathscr{C}}(n)$.

We also define a class of example distributions as a set of families $D = \{D_n\}_{n\in\mathbb{N}}$ of distributions, where D_n is a distribution on $\{0,1\}^n$. For any class $\mathcal{D}$ of example distributions and $n \in \mathbb{N}$, we use the notation $\mathcal{D}_n$ to represent $\mathcal{D}_n = \{D_n : \{D_m\}_{m\in\mathbb{N}} \in \mathcal{D}\}$.

Here, the PAC learning and agnostic learning models are defined as follows.

Definition 2 (PAC learning and agnostic learning [Val84, KSS94])**.** *Let $\mathscr{C}$ be a concept class and $\mathcal{D}$ be a class of example distributions. We say that a randomized oracle machine L, referred to as an agnostic learner,* agnostically learns *$\mathscr{C}$ on $\mathcal{D}$ (with time complexity $t : \mathbb{N} \times (0,1] \to \mathbb{N}$ and sample complexity $m : \mathbb{N} \times (0,1] \to \mathbb{N}$) if L satisfies the following conditions:*

1. *L is given $n \in \mathbb{N}$ and an accuracy parameter $\epsilon \in (0,1]$ as the input and given access to an example oracle $EX(f,D)$ determined by a (possibly randomized[4]) target function $f : \{0,1\}^n \to \{0,1\}$ and an example distribution $D \in \mathcal{D}_n$.*

2. *For each access, $EX(f,D)$ returns an example of the form $(x, f(x))$, where x is selected identically and independently according to D.*

3. *For all $n \in \mathbb{N}$, $\epsilon \in (0,1]$, randomized target functions $f : \{0,1\}^n \to \{0,1\}$, and example distributions $D \in \mathcal{D}_n$, the learner L outputs a circuit $h : \{0,1\}^n \to \{0,1\}$ as a hypothesis that is ϵ-close to f under D with probability at least $2/3$, i.e., L satisfies the following condition:*

$$\Pr_{L,EX(f,D)}\left[L^{EX(f,D)}(n,\epsilon) \text{ outputs } h \text{ such that } \Pr_{f,x\leftarrow D}[h(x) \neq f(x)] \leq \mathsf{opt}_{\mathscr{C},f} + \epsilon\right] \geq 2/3,$$

where $\mathsf{opt}_{\mathscr{C},f} = \min_{c^*\in\mathscr{C}} \Pr_{f,x\leftarrow D}[c^*(x) \neq f(x)]$.

[4] In other words, we assume that each $f(x)$ is associated with some distribution D_x on $\{0,1\}$, and the outcome of $f(x)$ is selected according to D_x.

4. L halts in time $t(n,\epsilon)$ with access to the example oracle at most $m(n,\epsilon)$ times in each case.

A PAC learner L (with time complexity $t(n,\epsilon)$ and sample complexity $m(n,\epsilon)$) on $\mathcal{D}$ is defined as a randomized oracle machine L satisfying the above-mentioned conditions 1, 2, and 4, as well as condition 3, except that we only consider the case of $f \in \mathscr{C}$, i.e., $\mathsf{opt}_{\mathscr{C},f} = 0$ (instead of all randomized target functions).

We say that $\mathscr{C}$ is agnostic (resp. PAC) learnable in polynomial time on $\mathcal{D}$ if there is an agnostic (resp. PAC) learner for $\mathscr{C}$ on $\mathcal{D}$ with time complexity $t \leq \mathsf{poly}(n, \epsilon^{-1})$. In addition, we say that $\mathscr{C}$ is weakly learnable on $\mathcal{D}$ if there exists a PAC learner for $\mathscr{C}$ on $\mathcal{D}$ with some fixed accuracy parameter $\epsilon \leq 1/2 - 1/\mathsf{poly}(n)$ and time complexity $t(n) \leq \mathsf{poly}(n)$.

We may grant a learner oracle access to a target function f, referred to as a membership query.

For a function $s : \mathbb{N} \to \mathbb{N}$, we define a concept class[5] $\mathsf{SIZE}[s]$ of circuits by

$$\mathsf{SIZE}[s] = \{f : \{0,1\}^n \to \{0,1\} | n \in \mathbb{N} \text{ and } f \text{ is computable by an } s(n)\text{-size circuit}\}.$$

3.2 Average-Case Complexity Theory

We define a distributional problem as a pair of a language $L \subseteq \{0,1\}^*$ and a distribution $D = \{D_n\}_{n\in\mathbb{N}}$ on instances where D_n is a distribution on $\{0,1\}^n$. We say that a distributional problem (L, D) has an errorless heuristic algorithm A with failure probability at most $\epsilon : \mathbb{N} \to (0,1)$ if (1) A outputs $L(x)$ $(:= \mathbb{1}\{x \in L\})$ or $\perp$ (which represents "failure") for every $n \in \mathbb{N}$ and $x \in \mathrm{supp}(D_n)$ in $\mathsf{poly}(n)$ time, and (2) the failure probability that $A(x)$ outputs $\perp$ is bounded above by $\epsilon(n)$ for each $n \in \mathbb{N}$. We remark that an errorless heuristic algorithm never outputs an incorrect value $\neg L(x)$ for any $x \in \mathrm{supp}(D)$. We define a class AvgP of solvable distributional problems by

$$\mathsf{AvgP} = \{(L, D) : \forall p : \mathsf{poly}, \exists A : \text{an errorless heuristic algorithm for } (L, D) \text{ with error at most } 1/p(n)\}.$$

For a standard complexity class $\mathcal{C}$ (e.g., NP and PH), we also define its average-case extension $\mathsf{Dist}\mathcal{C}$ as $\{(L, D) : L \in \mathcal{C}, D \text{ is polynomial-time samplable}\}$, where we say that $D = \{D_n\}_{n\in\mathbb{N}}$ is polynomial-time samplable if there exists a randomized sampling algorithm S such that $S(1^n) \equiv D_n$ for each $n \in \mathbb{N}$. Further details on the background can be found in a survey [BT06a] on average-case complexity theory.

3.3 RRHS-Refutation

We remark that, for a randomized function $f : \{0,1\}^n \to \{0,1\}$, a concept class $\mathscr{C}$, and a distribution D on $\{0,1\}^n$, we define a correlation $\mathrm{Cor}_D(f, \mathscr{C}) \in [-1, 1]$ between f and $\mathscr{C}$ with respect to D by

$$\mathrm{Cor}_D(f, \mathscr{C}) := \max_{c\in\mathscr{C}_n} \mathrm{E}_{f,x\leftarrow D}\left[(-1)^{f(x)} \cdot (-1)^{c(x)}\right] = 2 \cdot \max_{c\in\mathscr{C}_n} \Pr_{f,x\leftarrow D}[f(x) = c(x)] - 1.$$

The following is the formal description of correlative RRHS-refutation.

Definition 3 (Correlative RRHS-refutation)**.** *Let $\mathscr{C}$ be a concept class, $m : \mathbb{N} \times (0,1] \to \mathbb{N}$ be a function, and $\mathcal{D}$ be a class of example distributions. We say that a randomized algorithm A correlatively random-right-hand-side-refutes (correlatively RRHS-refutes) $\mathscr{C}$ with sample complexity m on $\mathcal{D}$ if for any $n \in \mathbb{N}$, $\epsilon \in (0,1]$, and example distribution $D_n \in \mathcal{D}_n$, A satisfies the following conditions: on input $n, \in \mathbb{N}$, $\epsilon \in (0,1)$, and $m := m(n,\epsilon)$ samples $S = \left((x^{(i)}, b^{(i)})\right)_{i=1}^m$, where $x^{(i)} \in \{0,1\}^n$ and $b^{(i)} \in \{0,1\}$ for each $i \in [m]$,*

[5] $\mathsf{SIZE}[n^2]$ is regarded as a complete problem for learning in the following sense: if $\mathsf{SIZE}[n^2]$ is agnostic (resp. PAC) learnable iff all polynomially evaluated classes are agnostic (resp. PAC) learnable by the simple padding argument.

1. *Soundness: if the samples S are selected identically and independently according to $EX(f, D_n)$ for a randomized function f such that $\mathrm{Cor}_{D_n}(f, \mathscr{C}) \geq \epsilon$, then*

$$\Pr_{S,A}[A(n, \epsilon, S) \text{ outputs "correlative"}] \geq 2/3;$$

2. *Completeness: if the samples S are selected identically and independently according to $EX(f_R, D_n)$ for a truly random function f_R (i.e., each $b^{(i)}$ is selected uniformly at random), then*

$$\Pr_{S,A}[A(n, \epsilon, S) \text{ outputs "random"}] \geq 2/3.$$

We also say that $\mathscr{C}$ is correlatively RRHS-refutable with sample complexity m on $\mathcal{D}$ if there exists a randomized algorithm that correlatively RRHS-refutes $\mathscr{C}$ with sample complexity m on $\mathcal{D}$.

Theorem 5 ([KL18])**.** *Let $\mathscr{C}$ be a concept class, and let $\mathcal{D}$ be a class of example distributions. If $\mathscr{C}$ is correlatively RRHS-refutable on $\mathcal{D}$ with $m(n, \epsilon)$ samples in time $T(n, \epsilon)$, then $\mathscr{C}$ is agnostic learnable on $\mathcal{D}$ with sample complexity $O(\frac{m(n,\epsilon/2)^3}{\epsilon^2})$ and time complexity $O(T(n, \epsilon/2) \cdot \frac{m(n,\epsilon/2)^2}{\epsilon^2})$.*

3.4 GapMINKT

The approximation problem of computing the time-bounded Kolmogorov complexity is formally defined as follows.

Definition 4 (Gap_τMINKT)**.** *For a function $\tau : \mathbb{N} \to \mathbb{N}$, Gap_τMINKT is a promise problem (Π_Y, Π_N) defined as follows:*

$$\Pi_Y = \left\{(x, 1^s, 1^t) : K^t(x) \leq s\right\},$$
$$\Pi_N = \left\{(x, 1^s, 1^t) : K^{\tau(|x|+t)}(x) > s + \log \tau(|x| + t)\right\}.$$

Hirahara [Hir20] showed that the above problem is efficiently solvable if $\mathsf{DistNP} \subseteq \mathsf{AvgP}$.

Theorem 6 ([Hir20])**.** *If $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, then $\mathrm{Gap}_\tau\mathrm{MINKT} \in \mathsf{pr\text{-}P}$ for some polynomial τ.*

Every algorithm A that solves Gap_τMINKT yields the approximation algorithm $\mathsf{ApproxK}_\tau$ simply as follows. On input $x \in \{0,1\}^*$ and 1^t, where $t \in \mathbb{N}$, $\mathsf{ApproxK}_\tau$ outputs the minimum $s \in \mathbb{N}$ such that $A(x, 1^s, 1^t) = 1$. Since $(x, 1^{s-1}, 1^t)$ is not a YES instance and $(x, 1^s, 1^t)$ is not a NO instance for such s, the following lemma is easily verified.

Lemma 1. *If $\mathrm{Gap}_\tau\mathrm{MINKT} \in \mathsf{pr\text{-}P}$, then there exists an algorithm $\mathsf{ApproxK}_\tau$ that is given $(x, 1^t)$, where $x \in \{0,1\}^*, t \in \mathbb{N}$, and outputs an integer $s \in \mathbb{N}$ in polynomial time to satisfy*

$$K^{\tau(|x|+t)}(x) - \log \tau(|x| + t) \leq s \leq K^t(x).$$

3.5 Weak Symmetry of Information

We introduce the following powerful tool available in Heuristica.

Theorem 7 (Weak symmetry of information [Hir21])**.** *If $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, then there exist polynomials p_0 and p_w that, for any $n, m \in \mathbb{N}$, $t \geq p_0(nm)$, $\epsilon \in (0, 1]$, and $x \in \{0,1\}^n$, satisfy*

$$\Pr_{r \sim \{0,1\}^m}\left[K^t(x \circ r) \geq K^{p_w(t/\epsilon)}(x) + m - \log p_w(t/\epsilon)\right] \leq \epsilon.$$

In this paper, we use the notations p_0 and p_w to refer to the polynomials in Theorem 7.

4 Agnostic Learning in Heuristica

In this section, we construct the agnostic learner based on $\mathsf{DistNP} \subseteq \mathsf{AvgP}$ and prove Theorem 1.

4.1 Agnostic Learning on Shallow Sampling-Depth Distributions

We present the construction of the agnostic learner and show the correctness on distributions that have shallow sampling-depth functions. We remark that sampling-depth functions of a distribution and a class of distributions are defined as follows.

Definition 5 (Sampling-depth functions)**.** *Let $t, t' \in \mathbb{N}$ such that $t' > t$. For a family of distributions D, we define a (t,t')-sampling-depth function $sd_D^{t,t'} = \{sd_{D,n}^{t,t'}\}_{n\in\mathbb{N}}$, where $sd_{D,n}^{t,t'} \colon \mathbb{N} \to \mathbb{R}_{\geq 0}$ by*

$$sd_{D,n}^{t,t'}(m) = \mathrm{E}_{X \leftarrow D_n^m}\left[K^t(X) - K^{t'}(X)\right].$$

We also extend the above-mentioned notion to a class of distributions. For a class $\mathcal{D}$ of families of distributions, we define a (t,t')-sampling-depth function $sd_{\mathcal{D}}^{t,t'} = \{sd_{\mathcal{D},n}^{t,t'}\}_{n\in\mathbb{N}}$, where $sd_{\mathcal{D},n}^{t,t'} \colon \mathbb{N} \to \mathbb{R}_{\geq 0}$ by

$$sd_{\mathcal{D},n}^{t,t'}(m) = \max_{D\in\mathcal{D}} sd_{D,n}^{t,t'}(m).$$

Our technical theorem is stated as follows.

Theorem 8. *For any polynomial $\tau \colon \mathbb{N} \to \mathbb{N}$, there exist polynomials $p_\tau(n,m,t)$ and $p'_\tau(n,m,t)$ satisfying the following. If $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, then there exists a learner L that agnostically learns $\mathscr{C}$ on $\mathcal{D}$ in time $\mathsf{poly}(n, m(n,\epsilon/2), t(n,\epsilon/2), \epsilon^{-1})$ with sample complexity $O(\epsilon^{-2}\cdot m(n,\epsilon/2)^3)$, where $m, t \colon \mathbb{N}\times(0,1] \to \mathbb{N}$ are arbitrary functions satisfying the following conditions: for all sufficiently large n and for all $\epsilon \in (0,1]$,*

$$\begin{aligned} t(n,\epsilon) &\geq p_0(nm(n,\epsilon)^2), \text{ and} \\ m(n,\epsilon) &> \frac{8}{\epsilon^2}\left(n + \ell_{\mathscr{C}}(n) + 6sd_{\mathcal{D},n}^{t(n,\epsilon),p_\tau(n,m(n,\epsilon),t(n,\epsilon))}(m(n,\epsilon)) + \log p'_\tau(n,m(n,\epsilon),t(n,\epsilon))\right). \end{aligned}$$

Proof. Let $m := m(n,\epsilon)$ and $t := t(n,\epsilon)$. First, we specify the polynomials p_τ and p'_τ. Fix $x^{(1)},\dots,x^{(m)} \in \{0,1\}^n$ and $f \in \mathscr{C}_n$ arbitrarily. Let $X = x^{(1)}\circ\cdots\circ x^{(m)}$. Then, we can compute $f(x^{(1)}),\dots,f(x^{(m)})$ in time $m\cdot\mathsf{poly}(n)$ from X, the representation of f, and the evaluation algorithm for $\mathscr{C}$ (where we use the assumption that $\ell_{\mathscr{C}}(n) \leq \mathsf{poly}(n)$ and $\mathscr{C}$ is polynomially evaluatable). For any $b \in \{0,1\}^m$ such that $|\{i \in [m] : b_i = f(x^{(i)})\}| \geq (1/2+\epsilon/4)m$, we define $e \in \{0,1\}^m$ by $e_i = b_i \oplus f(x^{(i)})$. Then, e is reconstructed from $H_2(1/2+\epsilon/4)m$ bits in time $\mathsf{poly}(n,m)$ by lexicographic indexing among binary strings of the same weight, where H_2 is the binary entropy function. Therefore, we can take a polynomial $t'(n,m,t)$ such that, for any sufficiently large n,

$$\begin{aligned} K^{t'(n,m,t)}(X\circ b) &\leq K^{\tau(nm+t)}(X) + \ell_{\mathscr{C}}(n) + n + H_2(1/2+\epsilon/4)m \\ &\leq K^{\tau(nm+t)}(X) + \ell_{\mathscr{C}}(n) + n + (1-\epsilon^2/8)\cdot m, \end{aligned} \tag{1}$$

where we applied the Taylor series of H_2 in a neighborhood of $1/2$, i.e., for any $\delta \in [-1/2, 1/2]$,

$$H_2(1/2+\delta) = 1 - \frac{1}{2\ln 2}\sum_{i=1}^{\infty}\frac{(2\delta)^{2i}}{i(2i-1)} \leq 1 - \frac{2}{\ln 2}\delta^2 \leq 1 - 2\delta^2.$$

Now, we define the polynomials p_τ and p'_τ by

$$p_\tau(n,m,t) = p_w(6\tau(nm + t'(n,m,t))), \text{ and}$$
$$p'_\tau(n,m,t) = p_w(6\tau(nm + t'(n,m,t)))\tau(nm + t'(n,m,t))\tau(nm + t)$$

Next, we construct a refutation algorithm R for $\mathscr{C}$ as follows. On input $n \in \mathbb{N}$, a set $S = \big((x^{(1)}, b^{(1)}), \ldots, (x^{(m)}, b^{(m)})\big)$ of samples, and $\epsilon \in (0,1]$, R computes t and $t' := t'(n,m,t)$, executes $s \leftarrow \mathsf{ApproxK}_\tau(X, 1^t)$ and $s' \leftarrow \mathsf{ApproxK}_\tau(X \circ b, 1^{t'})$ for $X = x^{(1)} \circ \cdots \circ x^{(m)}$ and $b = b^{(1)} \circ \cdots \circ b^{(m)}$, and finally outputs "correlative" if $s' - s \leq m + \ell_{\mathscr{C}}(n) + n + \log \tau(nm + t) - m\epsilon^2/8$ and outputs "random" otherwise.

We can easily verify that R halts in polynomial time in n, m, and t. We now verify the correctness of R. Let f denote a target randomized function for refutation.

In "correlative" cases, there exists a function $f^* \in \mathscr{C}_n$ such that

$$\Pr_{x \leftarrow D, f}[f(x) = f^*(x)] = \frac{1}{2} + \frac{Cor_D(f, \mathscr{C})}{2} \geq \frac{1}{2} + \frac{\epsilon}{2}.$$

According to the Hoeffding inequality, the probability that $|\{i \in [m] : b^{(i)} = f^*(x^{(i)})\}| < 1/2 + \epsilon/4$ holds is less than $\exp(-2m \cdot (\epsilon/4)^2) \leq \exp(-n \cdot 8/\epsilon^2 \cdot \epsilon^2/8) \leq 1/3$ over the choice of S for any sufficiently large $n \in \mathbb{N}$. In such cases, by Lemma 1 and inequality (1), we have

$$\begin{aligned} s' &\leq K^{t'(n,m,t)}(X \circ b) \\ &\leq K^{\tau(nm+t)}(X) + \ell_{\mathscr{C}}(n) + n + (1 - \epsilon^2/8) \cdot m \\ &\leq s + \log \tau(nm + t) + \ell_{\mathscr{C}}(n) + n + (1 - \epsilon^2/8) \cdot m, \end{aligned}$$

and

$$s' - s \leq m + \ell_{\mathscr{C}}(n) + n + \log \tau(nm + t) - m\epsilon^2/8.$$

Thus, $R(n, S, \epsilon)$ outputs "correlative" with a probability of at least $2/3$.

In "random" cases, b is selected uniformly at random from $\{0,1\}^m$. By the assumption that $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, $t \geq p_0(nm \cdot m)$, and the weak symmetry of information (Theorem 7), for any $X \in \{0,1\}^{nm}$, we have

$$\Pr_b\left[K^{\tau(nm+t'(n,m,t))}(X \circ b) \geq K^{p_w(6\tau(nm+t'(n,m,t)))}(X) + m - \log p_w(6\tau(nm + t'(n,m,t)))\right] \leq 1/6.$$

Let $D \in \mathcal{D}$ be an arbitrary example distribution. By Markov's inequality, we can show that

$$\Pr_X\left[K^t(X) - K^{p_\tau(n,m,t)}(X) > 6sd^{t,p_\tau(n,m,t)}_{\mathcal{D},n}(m)\right] \leq \frac{sd^{t,p_\tau(n,m,t)}_{D,n}(m)}{6sd^{t,p_\tau(n,m,t)}_{\mathcal{D},n}(m)} \leq \frac{1}{6}.$$

Thus, the following inequality holds with a probability of at least $1 - (1/6 + 1/6) = 2/3$:

$$\begin{aligned} s' &\geq K^{\tau(nm+t'(n,m,t))}(X \circ b) - \log \tau(nm + t'(n,m,t)) \\ &\geq K^{p_w(6\tau(nm+t'(n,m,t)))}(X) + m - \log p_w(6\tau(nm + t'(n,m,t)))\tau(nm + t'(n,m,t)) \\ &\geq K^{p_\tau(n,m,t)}(X) + m - \log p'_\tau(n,m,t) + \log \tau(nm + t) \\ &\geq K^t(X) - (K^t(X) - K^{p_\tau(n,m,t)}(X)) + m - \log p'_\tau(n,m,t) + \log \tau(nm + t) \\ &\geq s - (K^t(X) - K^{p_\tau(n,m,t)}(X)) + m - \log p'_\tau(n,m,t) + \log \tau(nm + t) \\ &\geq s - 6sd^{t,p_\tau(n,m,t)}_{\mathcal{D},n}(m) + m - \log p'_\tau(n,m,t) + \log \tau(nm + t). \end{aligned}$$

By arranging the above, we get

$$s' - s \geq m - 6sd_{\mathcal{D},n}^{t,p_\tau(n,m,t)}(m) - \log p'_\tau(n,m,t) + \log \tau(nm+t).$$

By the assumption that

$$m > \frac{8}{\epsilon^2}\left(n + \ell_{\mathscr{C}}(n) + 6sd_{\mathcal{D},n}^{t,p_\tau(n,m,t)}(m) + \log p'_\tau(n,m,t)\right),$$

we have

$$\begin{aligned}
&\left(m - 6sd_{\mathcal{D},n}^{t,p_\tau(n,m,t)}(m) - \log p'_\tau(n,m,t) + \log \tau(nm+t)\right) - \left(m + \ell_{\mathscr{C}}(n) + n + \log \tau(nm+t) - m\epsilon^2/8\right) \\
&= m\epsilon^2/8 - \left(n + \ell_{\mathscr{C}}(n) + 6sd_{\mathcal{D},n}^{t,p_\tau(n,m,t)}(m) + \log p'_\tau(n,m,t)\right) \\
&> 0.
\end{aligned}$$

Thus, $s' - s > m + \ell_{\mathscr{C}}(n) + n + \log \tau(nm+t) - m\epsilon^2/8$ holds, and R outputs "random" in such cases. Therefore, $R(n, S, \epsilon)$ outputs "random" with a probability of at least $2/3$

By the above-mentioned argument, R correlatively RRHS-refutes $\mathscr{C}$ on $\mathcal{D}$ in time $\mathsf{poly}(n,m,t)$ with m samples. Thus, by Theorem 5, we conclude that $\mathscr{C}$ is agnostic learnable on $\mathcal{D}$ in time

$$O\left(\mathsf{poly}(n, m(n,\epsilon/2), t(n,\epsilon/2)) \cdot \frac{m(n,\epsilon/2)^2}{\epsilon^2}\right) = \mathsf{poly}(n, m(n,\epsilon/2), t(n,\epsilon/2), \epsilon^{-1}),$$

with $O(\frac{m(n,\epsilon/2)^3}{\epsilon^2})$ samples. □

4.2 Sampling-Depth of P/poly-Samplable Distributions

Next, we observe that P/poly-samplable distributions have a logarithmically small sampling depth. Then, we use Theorem 8 to establish the agnostic learnability on P/poly-samplable distributions.

Definition 6 (P/poly-samplable distributions)**.** *For functions $t, a\colon \mathbb{N} \to \mathbb{N}$, we define a class $\mathsf{Samp}[t(n)]/a(n)$ of example distributions as follows: $D_n \in \mathsf{Samp}[t(n)]/a(n)_n$ iff there exist a randomized TM M and advice $z_n \in \{0,1\}^{a(n)-|M|}$ such that D_n is statistically identical to the distribution of outputs of $M(1^n; z_n)$ in $t(n)$ steps.*

To analyze the sampling depth of $\mathsf{Samp}[t(n)]/a(n)$, we introduce the following useful lemmas.

Lemma 4 ([Hir21, Corollary 9.8])**.** *If $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, then there exists a polynomial $p\colon \mathbb{N}\times\mathbb{N} \to \mathbb{N}$ such that for any $t, a\colon \mathbb{N} \to \mathbb{N}$, $n, m \in \mathbb{N}$, $D_n \in \mathsf{Samp}[t(n)]/a(n)_n$, and $x \in \mathrm{supp}(D_n^m)$,*

$$K^{p(t(n),m)}(x) \leq -\log D_n^m(x) + O(\log m) + O(\log t(n)) + a(n).$$

The following holds by the noiseless coding theorem.

Lemma 5 (cf. [LV19, Theorem 8.1.1])**.** *For any distribution D, $\mathrm{E}_{x\leftarrow D}[K(x)] \geq H(D)$.*

Now, we show the upper bound on the sampling depth of $\mathsf{Samp}[t(n)]/a(n)$.

Lemma 6. *If $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, then there exists a polynomial $p'_0\colon \mathbb{N}\times\mathbb{N} \to \mathbb{N}$ such that for any $t, a\colon \mathbb{N} \to \mathbb{N}$, $n, m \in \mathbb{N}$, the following expression holds: for all $t' \geq p'_0(t(n), m)$,*

$$sd_{\mathsf{Samp}[t]/a,n}^{t',\infty}(m) \leq O(\log m + \log t(n)) + a(n).$$

In this paper, we use the notation p'_0 to refer to the polynomial in Lemma 6.

Proof. Fix $D \in \mathsf{Samp}[t(n)]/a(n)$ arbitrarily. Let p'_0 denote the polynomial p in Lemma 4. Assuming that $\mathsf{DistNP} \subseteq \mathsf{AvgP}$ and Lemma 4, for each $m \in \mathbb{N}$, we have

$$\begin{aligned}
\mathrm{E}_{x \leftarrow D_n^m}[K^{t'}(x)] &\le \mathrm{E}_{x \leftarrow D_n^m}[K^{p'_0(t(n),m)}(x)] \\
&\le \mathrm{E}_{x \leftarrow D_n^m}[-\log D_n^m(x)] + O(\log m + \log t(n)) + a(n) \\
&\le H(D_n^m) + O(\log m + \log t(n)) + a(n) \\
&\le \mathrm{E}_{x \leftarrow D_n^m}[K(x)] + O(\log m + \log t(n)) + a(n).
\end{aligned}$$

Thus, we conclude that

$$\begin{aligned}
sd_{D,n}^{t',\infty}(m) &= \mathrm{E}_{x \leftarrow D_n^m}[K^{t'}(x)] - \mathrm{E}_{x \leftarrow D_n^m}[K(x)] \\
&\le O(\log m + \log t(n)) + a(n).
\end{aligned}$$

□

Theorem 8 and Lemma 6 imply the following corollary, which is the formal statement of Theorem 1.

Corollary 1. *If* $\mathsf{DistNP} \subseteq \mathsf{AvgP}$*, then for any polynomials* $s, t_s, a_s \colon \mathbb{N} \to \mathbb{N}$*,* $\mathsf{SIZE}[s(n)]$ *is agnostic learnable on* $\mathsf{Samp}[t_s(n)]/a_s(n)$ *in polynomial time with sample complexity* $O(\epsilon^{-8+\Delta}(n + s(n) + a_s(n))^{3+\Delta})$*, where* $\Delta > 0$ *is an arbitrary small constant.*

We remark that the time complexity t_s for the sampling algorithms above affects only the time complexity of the agnostic learner.

Proof. Let $\Delta > 0$ be an arbitrary small constant. By the assumption that $\mathsf{DistNP} \subseteq \mathsf{AvgP}$ and Theorem 6, there exists a polynomial τ such that $\mathrm{Gap}_\tau\mathrm{MINKT} \in \mathsf{pr\text{-}P}$. Thus, we can apply Theorem 8.

We define the functions $m, t \colon \mathbb{N} \times (0,1] \to \mathbb{N}$ by

$$\begin{aligned}
m(n,\epsilon) &= (\epsilon^{-2} \log \epsilon^{-1} \cdot (n + s(n) \log s(n) + a_s(n)))^{1+\Delta}, \text{ and} \\
t(n,\epsilon) &= \max\{\lceil p_0(n \cdot m(n,\epsilon)^2) \rceil, \lceil p'_0(t_s(n), m(n,\epsilon)) \rceil\}.
\end{aligned}$$

Obviously, $t(n,\epsilon) > p_0(n \cdot m(n,\epsilon)^2)$ and $t(n,\epsilon) > p'_0(t_s(n), m(n,\epsilon))$ hold. By Lemma 6, we get

$$\begin{aligned}
sd_{\mathsf{Samp}[t_s]/a_s,n}^{t(n,\epsilon),p_\tau(n,m(n,\epsilon),t(n,\epsilon))}(m(n,\epsilon)) &\le sd_{\mathsf{Samp}[t_s]/a_s,n}^{t(n,\epsilon),\infty}(m(n,\epsilon)) \\
&\le O(\log m(n,\epsilon) + \log t_s(n)) + a_s(n).
\end{aligned}$$

It is easily verified that $t(n,\epsilon) \le \mathsf{poly}(n, s(n), t_s(n), a_s(n), \epsilon^{-1}) = \mathsf{poly}(n, \epsilon^{-1})$ and

$$\log p'_\tau(n, m(n,\epsilon), t(n,\epsilon)) \le O(\log n + \log \epsilon^{-1} + \log m(n,\epsilon)).$$

Therefore, for any sufficiently large $n \in \mathbb{N}$, we have

$$\begin{aligned}
&\frac{8}{\epsilon^2}\left(n + O(s(n)\log s(n)) + 6sd_{\mathsf{Samp}[t_s]/a_s,n}^{t(n,\epsilon),p_\tau(n,m(n,\epsilon),t(n,\epsilon))}(m(n,\epsilon)) + \log p'_\tau(n, m(n,\epsilon), t(n,\epsilon))\right) \\
&\le \epsilon^{-2} \cdot O(n + s(n)\log s(n) + \log \epsilon^{-1} + a_s(n) + \log m(n,\epsilon)) \\
&\le O(\epsilon^{-2} \log \epsilon^{-1}(n + s(n)\log s(n) + a_s(n))) \\
&\le m(n,\epsilon).
\end{aligned}$$

Thus, by Theorem 8, we conclude that $\mathsf{SIZE}[s(n)]$ is agnostic learnable in time

$$\mathsf{poly}(n, m(n, \epsilon/2), t(n, \epsilon/2), \epsilon^{-1}) = \mathsf{poly}(n, s(n), \epsilon^{-1}, t_s(n), a_s(n)) = \mathsf{poly}(n, \epsilon^{-1}).$$

The sample complexity is at most $O(\epsilon^{-2} m(n, \epsilon/2)^3)$, which is bounded above by $O(\epsilon^{-8+\Delta'}(n+s(n)+a_s(n))^{3+\Delta'})$ for an arbitrary small constant $\Delta' > 0$ by selecting a sufficiently small Δ compared to Δ' in the above-mentioned argument. □

Remark. In fact, it is not clear whether several techniques (e.g., the weak symmetry of information) developed in [Hir20, Hir21] can be relativized when we only assume that $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, owing to the pseudorandom generator construction presented in [BFP05]. However, all of them can be relativized under the stronger assumption that $\mathsf{Dist}\Sigma_2^p \subseteq \mathsf{AvgP}$ (refer to Appendix A). Thus, Theorem 8 and all the results in this section can be also relativized under the assumption that $\mathsf{Dist}\Sigma_2^p \subseteq \mathsf{AvgP}$. Furthermore, when we restrict the target to the efficient agnostic learning on $\mathsf{P/poly}$-samplable distributions (i.e., Theorem 1), we can obtain the same learnability result by using only relativized techniques from the following observations. First, the same upper bound of the sampling-depth function in Lemma 6 is obtained by applying the encoding developed in [AF09, AGvM+18] with additional random strings, and such additional random strings are available for learners. Second, the same upper bound of the sampling-depth function in Lemma 6 holds for $t' = \infty$; in this case, we can apply the symmetry of information for resource-unbounded Kolmogorov complexity instead of the weak symmetry of information in Theorem 8. Third, since the upper bound in Lemma 6 is logarithmically small, the algorithm for GapMINKT in [Hir18] with a worse approximation factor is sufficient, which can be relativized.

5 Switching Lemma on General Domains

In this section, we extend the switching lemma of a binary domain to general domains. Our proof mainly follows the proof presented by Razborov [Raz93].

We remark that, for $p \in [0, 1]$ and a set V of variables on alphabets Σ, we define a p-random restriction $\rho \colon V \to \Sigma \cup \{*\}$ by the following procedure. First, we select a random subset $S \subseteq V$ of size $\lfloor p|V| \rfloor$ uniformly at random. Then, we set $\rho(x) = *$ for $x \in S$ and assign a random value $\rho(x) \leftarrow_u \Sigma$ for each $x \in V \setminus S$.

Lemma 2. *For $m \in \mathbb{N}$, let $\Sigma_1, \ldots, \Sigma_m$ be finite sets of alphabets, and let $V_1, \ldots, V_m$ be disjoint sets of variables, where each variable in V_i takes a value in Σ_i. For each $i \in [m]$, let ρ_i be a p_i-random restriction to V_i, where $p_i \in [0, 1]$. Then, for any t-DNF ϕ on the variables in $V_1 \cup \ldots \cup V_m$ and $k \in \mathbb{N}$, we have*

$$\Pr_{\rho_1, \ldots, \rho_m}[\phi|_{\rho_1 \ldots \rho_m} \text{ is not expressed as } k\text{-CNF}] \leq O\left(mt \cdot \max_{i \in [m]} p_i |\Sigma_i|^2\right)^k.$$

Proof. Each literal in ϕ of the form $(x \neq a)$ (for some $a \in \Sigma_i$) is expressed as $\bigvee_{b \in \Sigma_i : b \neq a} (x = b)$. Note that if we apply this transformation to all literals of the form $(x \neq a)$ in ϕ and expand them to obtain a DNF formula, these operations do not change the width of the original DNF ϕ. Thus, without loss of generality, we can assume that ϕ does not contain any literal of the form $(x \neq a)$.

For each $i \in [m]$, let $M_i = |\Sigma_i|$, $N_i = |V_i|$, and $n_i = \lfloor p_i N_i \rfloor$. To prove the lemma, we assume that $\phi|_{\rho_1 \ldots \rho_m}$ is not expressed as k-CNF and show that $\rho = \rho_1 \cdots \rho_m$ has a short description for estimating the number of such restrictions.

We can select a partial assignment π to $V_1 \cup \ldots \cup V_m$ of size at least $k + 1$ such that $\phi|_{\rho\pi} \equiv 0$, but for any proper subrestriction π' of π, $\phi|_{\rho\pi'} \not\equiv 0$ (otherwise, $\phi|_\rho$ must be expressed as k-CNF).

We also select subrestrictions π_j of π and restrictions σ_j inductively on $j \le s\,(\le k)$ by the following procedure. Assume that $(\pi_1, \sigma_1), \ldots, (\pi_{i-1}, \sigma_{j-1})$ have been determined, and $\pi_1 \cdots \pi_{j-1} \not\equiv \pi$; if not, we complete the procedure. Since $\pi_1 \cdots \pi_{j-1}$ is a proper subrestriction of π, we have $\phi|_{\rho\pi_1 \cdots \pi_{j-1}} \not\equiv 0$, and we can select the first term τ_j (in some fixed order) such that the value of τ_j is not determined by $\rho\pi_1 \cdots \pi_{j-1}$. Since $\tau_j|_{\rho\pi} \equiv 0$, there must exist a set S_j of variables that are contained in τ_j, unassigned by $\pi_1 \cdots \pi_{j-1}$ but assigned by π. We define σ_j by a partial assignment to S_j, which is consistent with the literals in τ_j. We also define π_j by the corresponding subrestriction of π to S_j. This procedure is repeated until $\pi_1 \cdots \pi_j \equiv \pi$ holds; let s denote the index j at the end. For convenience, we trim S_s (and π_s, σ_s correspondingly) in some arbitrary manner to satisfy $k = |S_1 \cup \cdots \cup S_s|$.

For each $j \in [s]$, let P_j denote the set of indices in $[t]$, which indicates the position of the variables in S_i among the literals in τ_j, and let Q_j denote $Q_j = (\pi_j(v_1), \ldots, \pi_j(v_{|P_j|}))$, where $v_{j'}$ is the j'-th variable indicated by P_j. For each $i \in [m]$, let k_i be the number of variables in V_i that are assigned by $\sigma_1 \ldots \sigma_s$, i.e., we have $k = \sum_i k_i$.

We claim that ρ can be reconstructed from the composite restriction $\rho' = \rho\sigma_1 \cdots \sigma_s$, $P_1, \ldots, P_s$, and $Q_1, \ldots, Q_s$ by the following procedure: (0) let $j = 1$; (1) find the first term not to become 0 by ρ', which must be τ_j by the construction; (2) obtain σ_j and π_j from ρ', P_j, and Q_j; (3) let $\rho' := \rho\pi_1 \ldots \pi_j\sigma_{j+1} \ldots \sigma_m$ and $j := j + 1$, and repeat (1) and (2) to obtain σ_j and π_j; (4) repeat (3) until all of $\sigma_1, \ldots, \sigma_s$ are obtained; then, ρ can be reconstructed from ρ' and $\sigma_1, \ldots, \sigma_s$.

Therefore, ρ is represented by $P_1, \ldots, P_s, Q_1, \ldots, Q_s$, and the composite restriction ρ' that has $(n_i - k_i)$ $*$s on V_i for each $i \in [m]$. For each choice of $k_1, \ldots, k_m$ such that $k = \sum_i k_i$ and each $i \in [m]$, the possible choice of P_i is at most t^{k_i}, and the possible choice of Q_i is at most $M_i^{k_i}$. Thus, the possible number of such expressions is at most

$$C \cdot \sum_{k_i : k = \sum_i k_i} \prod_{i \in [m]} \binom{N_i}{n_i - k_i} \cdot M_i^{N_i - n_i + k_i} \cdot t^{k_i} \cdot M_i^{k_i} = C \cdot \sum_{k_i : k = \sum_i k_i} t^k \cdot \prod_{i \in [m]} \binom{N_i}{n_i - k_i} \cdot M_i^{N_i - n_i + k_i} \cdot M_i^{k_i},$$

for some absolute constant C.

If $\max_{i \in [m]} n_i/N_i \ge 1/2$, then the lemma holds trivially because $2 \max_{i \in [m]} p_i \ge 1$. Therefore, we can assume that $n_i/N_i < 1/2$, i.e., $n_i < N_i/2$ for each $i \in [m]$. Then, we can establish the upper bound on the probability as follows:

$$\begin{aligned}
\Pr_{\rho_1, \ldots, \rho_m}[\phi|_{\rho_1 \ldots \rho_m} \text{ is not expressed as } k\text{-CNF }] &\le C \cdot \sum_{k_i : k = \sum_i k_i} t^k \cdot \prod_{i \in [m]} \frac{\binom{N_i}{n_i - k_i} \cdot M_i^{N_i - n_i + 2k_i}}{\binom{N_i}{n_i} \cdot M_i^{N_i - n_i}} \\
&\le C \cdot \sum_{k_i : k = \sum_i k_i} t^k \cdot \prod_{i \in [m]} \frac{n_i^{k_i}}{(N_i - n_i)^{k_i}} M_i^{2k_i} \\
&\le C \cdot \sum_{k_i : k = \sum_i k_i} t^k \cdot \max_{i \in [m]} \left(\frac{n_i M_i^2}{N_i - n_i} \right)^k \\
&\le C \cdot (mt)^k \cdot \max_{i \in [m]} \left(\frac{n_i M_i^2}{N_i - n_i} \right)^k \\
&\le C \cdot (mt)^k \cdot \max_{i \in [m]} \left(\frac{2n_i M_i^2}{N_i} \right)^k \\
&= O\left(mt \cdot \max_{i \in [m]} \frac{n_i M_i^2}{N_i} \right)^k
\end{aligned}$$

$$= O\left(mt \cdot \max_{i\in[m]} p_i|\Sigma_i|^2\right)^k.$$

□

The above-mentioned lemma implies the following by considering the negation of a given CNF formula.

Lemma 7. *For $m \in \mathbb{N}$, let $\Sigma_1, \ldots, \Sigma_m$ be finite sets of alphabets, and let $V_1, \ldots, V_m$ be disjoint sets of variables, where each variable in V_i takes a value in Σ_i. For each $i \in [m]$, let ρ_i be a p_i-random restriction to V_i, where $p_i \in [0,1]$. Then, for any t-CNF ϕ on the variables in $V_1 \cup \ldots \cup V_m$ and $k \in \mathbb{N}$, we have*

$$\Pr_{\rho_1,\ldots,\rho_m}[\phi|_{\rho_1\ldots\rho_m} \text{ is not expressed as } k\text{-DNF }] \leq O\left(mt \cdot \max_{i\in[m]} p_i|\Sigma_i|^2\right)^k.$$

Now, we extend the above-mentioned results to constant-depth circuits on general domains. For any depth-d circuit, we number each layer from 0 (bottom) to d (top), where layer 0 consists of input gates and layer d consists of the topmost $\vee$- or $\wedge$-gate. Without loss of generality, we can assume that each depth-d circuit satisfies the following properties: (1) each input gate is a literal taking the form of either $(x = a)$ or $(x \neq a)$ for some alphabet a; (2) each layer (from 1 to d) contains either $\vee$-gates or $\wedge$-gates; and (3) the type of gate (i.e., $\vee$ or $\wedge$) alternates at adjacent layers. For any depth-d circuit, we define its width by the maximum number of literals (i.e., input gates) that are connected to the same gate and define its internal size by the total number of gates at layers $2, 3, \ldots d$. Then, our technical lemma is stated as follows.

Lemma 8. *For $m \in \mathbb{N}$, let $\Sigma_1, \ldots, \Sigma_m$ be finite sets of alphabets, and let $V_1, \ldots, V_m$ be disjoint sets of variables, where each variable in V_i takes a value in Σ_i. For each $i \in [m]$, let ρ_i be a p_i-random restriction to V_i, where $p_i \in [0,1]$. Then, for any depth-d circuit C on the variables in $V_1 \cup \ldots \cup V_m$ of width $\leq t$ and internal size $\leq c2^t$ (for some constant c), we have*

$$\Pr_{\rho_1,\ldots,\rho_m}[C|_{\rho_1\ldots\rho_m} \text{ is not a constant }] \leq O\left(mt \cdot \max_{i\in[m]} p_i^{\frac{1}{d}}|\Sigma_i|^2\right).$$

Proof. For each $i \in [d]$, let s_i be the number of gates at layer i.

First, we consider only the case where the following holds: for all $i \in [m]$,

$$\lfloor p_i N \rfloor \leq \lfloor p_i^{1/d} \lfloor p_i^{1/d} \cdots \lfloor p_i^{1/d} N \underbrace{\rfloor \cdots \rfloor}_{d-1 \text{ times}} \rfloor. \tag{2}$$

In this case, we can regard a p_i-restriction ρ_i as consecutive applications of $p_i^{1/d}$-random restrictions $\rho_i^{(1)}, \ldots, \rho_i^{(d-1)}$, and one remaining random restriction. For each $i \in [d-1]$, let $\rho^{(i)} \equiv \rho_1^{(i)} \ldots \rho_m^{(i)}$.

We assume that layer 1 consists of $\wedge$-gates (in the case of $\vee$-gates, we can show the lemma in the same manner). In this case, each gate in layer 2 is regarded as t-DNF; thus, we can apply Lemma 2 and show that all the gates at layer 2 are transformed into t-CNF with a probability of at least $1 - s_2 \cdot (c'mt \cdot \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2)^t$ for some absolute constant c'. If such an event occurs, then each $\vee$-gate at layer 2 collapses into its parent node. Thus, the depth decreases by 1. Since the resulting depth-$(d-1)$ circuit has width t, we can apply Lemma 7 and the same argument at layer 3. We repeat the same argument $(d-2)$ times for $\rho_i^{(1)}, \ldots, \rho_i^{(d-2)}$ at layers $2, \ldots, d$, respectively. Then,

the resulting circuit becomes a depth-2 circuit of width t (i.e., t-DNF or t-CNF) with a probability of at least

$$1 - (s_2 + s_3 + \ldots + s_d) \cdot (c'mt \cdot \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2)^t \geq 1 - (c2^t) \cdot (c'mt \cdot \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2)^t.$$

We apply Lemmas 2 and 7 and show that the resulting circuit becomes a constant by $\rho^{(d-1)}$ with a probability of at least $1 - c'mt \cdot \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2$. Without loss of generality, we can assume that $2c'mt \cdot \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2 < 1$; otherwise, the lemma holds trivially. Thus, we conclude that

$$\begin{aligned}
\Pr_{\rho_1,\ldots,\rho_m}[C|_{\rho_1\ldots\rho_m} \text{ is not a constant }] &\leq \Pr_{\rho^{(1)},\ldots,\rho^{(d-1)}}\left[C|_{\rho^{(1)}\ldots\rho^{(d-1)}} \text{ is not a constant }\right] \\
&\leq c(2c'mt \cdot \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2)^t + mt \cdot \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2 \\
&\leq 2cc'mt \cdot \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2 + mt \cdot \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2 \\
&= O\left(mt \cdot \max_{i\in[m]} p_i^{\frac{1}{d}}|\Sigma_i|^2\right).
\end{aligned}$$

Next, we consider the case where (2) does not hold for some $i \in [m]$. We can assume that $p_i^{1/d} < 1/4$; otherwise, we have $1/4 \leq p_i^{1/d} \leq \max_{i\in[m]} p_i^{1/d}|\Sigma_i|^2$, and the lemma holds trivially. In this case, we can show that

$$\begin{aligned}
p_iN &\geq \lfloor p_iN \rfloor \\
&> \lfloor\lfloor\cdots\lfloor p_i^{1/d}N\rfloor\cdots\rfloor\rfloor \\
&\geq \lfloor\lfloor\cdots\lfloor p_i^{1/d}(p_i^{1/d}N - 1)\rfloor\cdots\rfloor\rfloor \\
&\geq p_i^{(d-1)/d}N - (1 + p_i^{1/d} + p_i^{2/d} + \cdots + p_i^{(d-2)/d}) \\
&\geq p_i^{(d-1)/d}N - 2.
\end{aligned}$$

By rearranging the above, we have

$$\frac{2}{p_i^{(d-1)/d}N} \geq 1 - p_i^{1/d} > \frac{3}{4},$$

and

$$p_iN = p_i^{1/d} \cdot p_i^{(d-1)/d}N < \frac{1}{4} \cdot \frac{8}{3} = \frac{2}{3}.$$

Therefore, $\lfloor p_iN \rfloor = 0$ holds, and all the variables in V_i are fully determined by ρ_i. Thus, we can ignore such i in the argument above. □

6 Oracle Separation: UP ∩ coUP and Distributional PH

We improve the oracle separation in [Imp11] by applying Lemma 8. In Sections 6.1– 6.3, we present the following theorem (i.e., the first item of Theorem 3). In fact, the second item of Theorem 3 is shown in a similar way by changing the parameters (we will discuss this in Section 6.4).

Theorem 9. *For any function $\epsilon(n)$ such that $\omega(1) \leq \epsilon(n) \leq n/\omega(\log^2 n)$, there exists an oracle $\mathcal{O}_\epsilon$ satisfying (1) $\mathsf{DistPH}^{\mathcal{O}_\epsilon} \subseteq \mathsf{AvgP}^{\mathcal{O}_\epsilon}$ and (2) $\mathsf{UP}^{\mathcal{O}_\epsilon} \cap \mathsf{coUP}^{\mathcal{O}_\epsilon} \not\subseteq \mathsf{BPTIME}^{\mathcal{O}_\epsilon}[2^{O(\frac{n}{\epsilon(n)\log n})}]$.*

6.1 Construction of Random Oracle

Let $\epsilon : \mathbb{N} \to \mathbb{N}$ denote a parameter such that $\omega(1) \leq \epsilon(n) \leq n/\omega(\log^2 n)$.

Construction. $\mathcal{O}_\epsilon = \mathcal{V} + \mathcal{A}$, where each oracle is randomly selected by the following procedure:

1. Define functions t, p, ℓ, and $i_{\max}$ by

$$t(n) = 2^{\frac{n}{\epsilon(n)\cdot \log n}},\ p(n) = t(n)^{-\epsilon(n)^{1/2}},\ \ell(n) = t(n)^2,\ \text{and}\ i_{\max}(n) = \log\log t(n).$$

2. For each $n \in \mathbb{N}$, define a set $V_{n,0}$ of variables on alphabet Σ_n of size $\ell(n)$ by

$$V_{n,0} = \{F_x : x \in \{0,1\}^n\}.$$

   We assume that each alphabet in Σ_n has a binary representation of length at most $\lceil \log \ell(n) \rceil$.

3. For each $n \in \mathbb{N}$ and $i \in [i_{\max}(n) - 1]$, we inductively (on i) define a $p(n)$-random restriction $\rho_{n,i}$ to $V_{n,i-1}$, and we define a subset $V_{n,i} \subset V_{n,i-1}$ of variables by

$$V_{n,i} = \{v \in V_{n,i-1} : \rho_{n,i}(v) = *\}.$$

   We also define $\rho_{n,i_{\max}(n)}$ as a 0-random restriction (i.e., a full assignment) to $V_{n,i_{\max}(n)-1}$. Let $\rho_{n,i} \equiv \rho_{n,i_{\max}(n)}$ for $i \geq i_{\max}(n) + 1$. For simplicity, we may identify $\rho_{n,i}$ with a composite restriction $\rho_{n,1} \dots \rho_{n,i}$ to $V_{n,0}$ for each n and i.

4. Let $f = \{f_n\}_{n\in\mathbb{N}}$, where $f_n \colon \{0,1\}^n \to \Sigma_n$ is a random function defined by $f_n(x) = \rho_{n,i_{\max}(n)}(F_x)$.

5. Define $\mathcal{V}$ as follows:

$$\mathcal{V}(x,y) = \begin{cases} 1 & \text{if } y = f(x) \\ 0 & \text{otherwise.} \end{cases}$$

6. Define $\mathcal{A}$ as follows: On input $(\langle M, d\rangle, x, 1^{T^2})$, where M is an oracle machine, $d \in \mathbb{N}$, $x \in \{0,1\}^*$, and $T \in \mathbb{N}$, the oracle $\mathcal{A}$ returns the value in $\{0,1,?\}$ determined according to the following procedure:

   1: let $i := \log\log T$;
   2: construct a depth $d+2$ circuit C corresponding to the quantified formula

$$\exists w_1 \in \{0,1\}^{|x|} \forall w_2 \in \{0,1\}^{|x|}, \dots, Q_d w_d \in \{0,1\}^{|x|}, M^{\mathcal{O}}(x, w_1, w_2, \dots, w_d),$$

   where $Q_d = \exists$ if d is an odd number; otherwise, $Q_d = \forall$.
   First, we construct a depth d circuit that represents the above-mentioned quantified formula, where each leaf corresponds to $M^{\mathcal{O}}(x, w_1, w_2, \dots, w_d)$ for some $w_1, w_2, \dots, w_d$, where we truncate $w_1, w_2, \dots, w_d$ into a string of length T because we will execute M in only T steps. Then, we replace each leaf with a DNF formula of width T to obtain the circuit C, where each term corresponds to one possible choice of $\mathcal{V}$ such that $M^{\mathcal{V}+\mathcal{A}}(x, w_1, w_2, \dots, w_d)$ halts with an accepting state after execution in T steps. In other words, we consider each function $f' = \{f'_n\}_{n=1}^T$, where $f'_n \colon \{0,1\}^n \to \Sigma_n$, define an oracle $\mathcal{V}'$ in the same manner as $\mathcal{V}$, and execute $M^{\mathcal{V}'+\mathcal{A}}(x, w_1, w_2, \dots, w_d)$ in T steps. If M queries (x, y) to $\mathcal{V}'$, and the answer is 1 (resp. 0), then we add a literal $(F_x = y)$ (resp. $(F_x \neq y)$) to the corresponding term. Finally, we construct a circuit $C = \bigvee_{f'} C_{f'}$ on $V_{1,0}, \dots, V_{T,0}$.

By the construction described above, the above-mentioned quantified formula is satisfied with the execution of $M^{\mathcal{V}+\mathcal{A}}$ in T steps iff C returns 1 when it is restricted by $\rho_{1,i_{\max}(1)}, \ldots, \rho_{T,i_{\max}(T)}$. We can also easily verify that the width of C is at most T and the internal size of C is at most 2^{T+1}.
3: if $C|_{\rho_{1,i},\ldots,\rho_{T,i}} \equiv b$ for some $b \in \{0,1\}$, then **return** b; otherwise, **return** "?".

To verify that the above-mentioned $\mathcal{A}$ is not circular on recursive calls for $\mathcal{A}$, it is sufficient to show the following.

Lemma 9. *For each input, the value of $\mathcal{A}(\langle M, d\rangle, x, 1^{T^2})$ is determined by only $\rho_{n,j}$ for $n \leq T$ and $j \leq \log\log T$ $(= i)$.*

Proof. We show the lemma by induction on T. We consider the execution of $(\langle M, d\rangle, x, 1^{T^2})$. We remark that $\mathcal{A}$ first makes a depth $d+2$ circuit C based on M, and C is independent of the value of $\mathcal{V}$.

We assume that M makes some valid query $(\langle M', d'\rangle, x', 1^{T'^2})$ to $\mathcal{A}$ recursively on constructing C. Since the length of such a query is at most T, we have $T'^2 \leq T$. If we let $i' = \log\log T'$, then we have

$$i' = \log\log T' \leq \log\log T^{\frac{1}{2}} = \log\log T - 1.$$

By the induction hypothesis, the recursive answer of $\mathcal{A}$ is determined by only $\rho_{n,j}$ for $n \leq T'$ and $j \leq i-1$, and so is C. The lemma holds because the answer of $\mathcal{A}$ is determined by restricting C by $\rho_{n,j}$ for $n \leq T$ and $j \leq i$. □

Lemma 10. *For $\epsilon(n) = \omega(1)$, $V_{n,i_{\max}(n)-1} \neq \emptyset$ for sufficiently large n.*

Proof. Since $\epsilon(n) = \omega(1)$, we have $t(n) \leq 2^n$ and $i_{\max}(n) \leq \log n$ for sufficiently large n. Thus, for sufficiently large n, we have

$$p(n)^{i_{\max}(n)-1} \geq t(n)^{-\epsilon(n)^{1/2}\log n} = 2^{-\frac{n\log n}{\epsilon(n)^{1/2}\log n}} \geq 2^{-\frac{n}{\omega(1)}},$$

and $|V_{n,i_{\max}(n)}| = \Omega(p(n)^{i_{\max}(n)-1} \cdot 2^n) = 2^{\Omega(n)} > 0$. □

Note that we may omit the subscript ϵ from $\mathcal{O}_\epsilon$.

6.2 Worst-Case Hardness of UP ∩ coUP

Theorem 10. *For any function ϵ such that $\omega(1) \leq \epsilon(n) \leq n/\omega(\log^2 n)$, with probability 1 over the choice of $\mathcal{O}_\epsilon$, no randomized oracle machine can compute f within $t(n) = 2^{\frac{n}{\epsilon(n)\cdot\log n}}$ steps with a probability of at least $1 - 2^{-2n}$, where f is the random function selected in $\mathcal{O}_\epsilon$.*

Proof. We fix a randomized oracle machine A and input size $n \in \mathbb{N}$ arbitrarily. By the Borel–Cantelli lemma, union bound, and countability of randomized oracle machines, it is sufficient to show that for sufficiently large n,

$$\Pr_{\mathcal{O}}\left[\Pr_{A}\left[\forall x \in \{0,1\}^n, A^{\mathcal{O}}(x) \text{ outputs } f(x) \text{ within } t(n) \text{ steps}\right] \geq 1 - 2^{-n}\right] \leq n^{-\omega(1)}.$$

To show the above inequality, we prove the following and apply Markov's inequality:

$$P_{A,n} := \Pr_{A,\mathcal{O}}\left[\forall x \in \{0,1\}^n, A^{\mathcal{O}}(x) \text{ outputs } f(x) \text{ within } t(n) \text{ steps}\right] \leq n^{-\omega(1)}.$$

First, we fix random restrictions $\rho_{n',i}$ except for $\rho_{n,i_{\max}(n)}$ arbitrarily and use the notation ρ to denote the restriction. We remark that ρ determines $V_{n,i_{\max}(n)-1}$. Even under the condition on ρ, the value of $f_n(x)$ for each x such that $F_x \in V_{n,i_{\max}(n)-1}$ is selected from Σ_n uniformly at random (by $\rho_{n,i_{\max}(n)}$). Let x_ρ be the lexicographically first string $x_\rho \in \{0,1\}^n$ satisfying $F_{x_\rho} \in V_{n,i_{\max}(n)-1}$. Then, we also fix all remaining values of f_n except for $f_n(x_\rho)$, and let ρ' denote the restriction.

When we execute A in $t(n)$ steps, the length of the query made by A is at most $t(n)$. Thus, A can only access $\mathcal{A}(M, x, 1^{T^2})$ for $T \le t(n)^{1/2}$. For such T, we have

$$i = \log\log T \le \log\log t(n) - 1 = i_{\max}(n) - 1.$$

Therefore, by Lemma 9, the answers of $\mathcal{A}$ to queries made by A are determined only by ρ. Thus, for each random string for A, the queries made by $A(x_\rho)$ are also determined independently of the value of $f(x_\rho)$ unless A asks $(x_\rho, f(x_\rho))$ to $\mathcal{V}$. Since the value of $f(x_\rho)$ is selected uniformly at random under the condition on ρ and ρ', we have

$$\begin{aligned}
P_{A,n} &= \mathrm{E}_{\rho,\rho',r}\left[\Pr_{\mathcal{O}}\left[\forall x \in \{0,1\}^n,\ A^{\mathcal{O}}(x;r) \text{ outputs } f(x) \text{ within } t(n) \text{ steps} \middle| \rho, \rho'\right]\right] \\
&\le \mathrm{E}_{\rho,\rho',r}\left[\Pr_{\mathcal{O}}\left[A^{\mathcal{O}}(x_\rho;r) \text{ outputs } f(x_\rho) \text{ within } t(n) \text{ steps} \middle| \rho, \rho'\right]\right] \\
&\le \mathrm{E}_{\rho,\rho',r}\left[\frac{t(n)+1}{\ell(n)}\right] = O\left(\frac{t(n)}{t(n)^2}\right) = 2^{-\Omega(\frac{n}{\epsilon(n)\log n})} = n^{-\omega(1)},
\end{aligned}$$

where the last inequality follows from $\epsilon(n) \le \frac{n}{\omega(\log^2 n)}$. □

Corollary 2. *For any function $\epsilon(n)$ such that $\omega(1) \le \epsilon(n) \le n/\omega(\log^2 n)$, with probability 1 over the choice of $\mathcal{O}_{\epsilon'}$ for $\epsilon'(n) = \sqrt{\epsilon(n)}$, we have $\mathsf{UP}^{\mathcal{O}_{\epsilon'}} \cap \mathsf{coUP}^{\mathcal{O}_{\epsilon'}} \not\subseteq \mathsf{BPTIME}^{\mathcal{O}_{\epsilon'}}[2^{O(\frac{n}{\epsilon(n)\log n})}]$.*

Proof. Fix a random oracle $\mathcal{O} = \mathcal{O}_{\epsilon'}$ arbitrarily. For each alphabet $y \in \Sigma_n$ ($n \in \mathbb{N}$), we use the notation $\langle y \rangle$ to refer to its unique binary expression of length at most $\lceil \log \ell(n) \rceil$. We consider the following language $L^{\mathcal{O}}$:

$$L^{\mathcal{O}} = \{(x,i) : n \in \mathbb{N}, x \in \{0,1\}^n, i \in [n], \text{ and } \exists y \in \Sigma_n \text{ s.t. } f_n(x) = y \text{ and } \langle y \rangle_i = 1\}.$$

Obviously, $y := f_n(x)$ is a unique witness for both statements $\langle x, i\rangle \in L^{\mathcal{O}}$ and $\langle x, i\rangle \notin L^{\mathcal{O}}$ by verifying whether $\mathcal{V}(x,y) = 1$ and $y_i = 1$ hold. Thus, $L^{\mathcal{O}} \in \mathsf{UP}^{\mathcal{O}} \cap \mathsf{coUP}^{\mathcal{O}}$.

Suppose that $\mathsf{UP}^{\mathcal{O}} \cap \mathsf{coUP}^{\mathcal{O}} \subseteq \mathsf{BPTIME}^{\mathcal{O}}[2^{O(n/(\epsilon(n)\log n))}]$. Then, there exists a randomized oracle machine $A^{\mathcal{O}}$ that solves $L^{\mathcal{O}}$ in time $2^{an/(\epsilon(n)\log n)}$ with a probability of at least $2/3$ for some constant $a > 0$. Now, we can construct a randomized oracle machine $B^{\mathcal{O}}$ to compute f as follows. On input $x \in \{0,1\}^n$, $B^{\mathcal{O}}$ executes $b_i = A^{\mathcal{O}}(\langle x, i\rangle)$ for each $i \in [n]$, where $B^{\mathcal{O}}$ executes $A^{\mathcal{O}}$ $\mathsf{poly}(n)$ times and takes the majority of the answers to reduce the error probability of A from $1/3$ to $1/(n2^{2n})$.

By the union bound, the error probability of B is at most 2^{2n}. Let $n' = |\langle x, i\rangle|$ for $x \in \{0,1\}^n$ and $i \in [n]$. Then, we can assume that $n' \le 2n$ for sufficiently large n. Thus, the running time of B is bounded above by $\mathsf{poly}(n) \cdot 2^{2an/(\epsilon(2n)\log 2n)}$, which is less than $2^{n/(\epsilon(n)'\log n)}$ for sufficiently large n. Since such $\mathcal{O}$ contradicts the statement in Theorem 10, we conclude that the event that $\mathsf{UP}^{\mathcal{O}} \cap \mathsf{coUP}^{\mathcal{O}} \not\subseteq \mathsf{BPTIME}^{\mathcal{O}}[2^{O(\frac{n}{\epsilon(n)\log n})}]$ occurs with probability 1 over the choice of $\mathcal{O}$. □

6.3 Average-Case Easiness of PH

Theorem 11. *For any function $\omega(1) \le \epsilon(n) \le n/\omega(\log^2 n)$, the following event occurs with probability 1 over the choice of $\mathcal{O}_\epsilon$: for all triples of a polynomial-time oracle machine $M^?$, $d \in \mathbb{N}$, and a*

polynomial-time randomized oracle sampling machine $S^?$, there exists a deterministic polynomial-time errorless heuristic oracle machine with a failure probability of at most $n^{-\omega(1)}$ for the distributional Σ_d^p problem $(L_M^{\mathcal{O}}, D_S^{\mathcal{O}})$ determined as follows: $(D_S^{\mathcal{O}})_n \equiv S^{\mathcal{O}}(1^n)$ for each $n \in \mathbb{N}$ and

$$L_M^{\mathcal{O}} = \{x \in \{0,1\}^* : \exists w_1 \in \{0,1\}^{|x|} \forall w_2 \in \{0,1\}^{|x|}, \ldots, Q_d w_d \in \{0,1\}^{|x|}, M^{\mathcal{O}}(x, w_1, w_2, \ldots, w_d) = 1\},$$

where $Q_d = \exists$ if d is an odd number; otherwise, $Q_d = \forall$.

By the padding argument on the instance, the above-mentioned theorem implies the following.

Corollary 3. *For any function $\omega(1) \le \epsilon(n) \le n/\omega(\log^2 n)$, the event $\mathsf{DistPH}^{\mathcal{O}_\epsilon} \subseteq \mathsf{AvgP}^{\mathcal{O}_\epsilon}$ occurs with probability 1 over the choice of $\mathcal{O}_\epsilon$.*

Theorem 9 immediately follows from Corollaries 2 and 3.

Proof of Theorem 11. For each n, let T_n be the maximum value of n, the square of the time for $S^?$ to generate an instance of size n, and the square of the time to execute $M^?$ on instance size n. Let $i_n = \log\log T_n$.

Now, we construct an errorless heuristic scheme $B^{\mathcal{O}}$ that is given $x \in \{0,1\}^n$ as input and returns a value of $\mathcal{A}(\langle M, d\rangle, x, 1^{T_n^2})$. Remember that $B^{\mathcal{O}}(x) = L_M^{\mathcal{O}}(x)$ unless $\mathcal{A}(\langle M, d\rangle, x, 1^{T_n^2})$ outputs "?". Thus, we show the inequality

$$P_{n,M,S} := \Pr_{\mathcal{O},S}\left[\mathcal{A}(\langle M, d\rangle, x, 1^{T_n^2}) = \text{"?"} \text{ where } x \leftarrow S^{\mathcal{O}}(1^n)\right] \le n^{-\omega(1)}. \tag{3}$$

Then, by applying Markov's inequality, we have

$$\Pr_{\mathcal{O}}\left[\Pr_{S}\left[B^{\mathcal{O}}(x) = L_M^{\mathcal{O}}(x) \text{ where } x \leftarrow S(1^n)\right] > n^{-\omega(1)}\right] \le \frac{1}{n^2},$$

and the theorem follows from the Borel–Cantelli lemma and the countability of (M, d, S).

To show inequality (3), we first show that the instance $x \in \{0,1\}^n$ is determined by only $\rho_{n',j}$ for $n' \le T_n$ and $j \le i_n - 1$ with a probability of at least $1 - n^{-\omega(1)}$. Then, we will show that $\mathcal{A}(\langle M, d\rangle, x, 1^{T_n^2})$ returns $L_M^{\mathcal{O}}(x)$ (i.e., $\mathcal{A}(\langle M, d\rangle, x, 1^{T_n^2}) \neq$"?") with a probability of at least $1 - n^{-\omega(1)}$ under the condition that x is determined by only $\rho_{n',j}$ for $n' \le T_n$ and $j \le i_n - 1$.

Let T_n^S be the time bound for S to generate an instance of size n. Since $T_n^S \le T_n^{1/2}$, the answers of $\mathcal{A}$ to queries made by $S(1^n)$ are determined by only $\rho_{n',j}$ for $n' \le T_n^{1/2}$ and $j \le i_n - 2$. Under the condition on restrictions $\rho_{n',j}$ for $n' \le T_n^{1/2}$ and $j \le i_n - 2$, the value of $x \leftarrow S^{\mathcal{O}}(1^n)$ is determined by only $\rho_{n',j}$ for $n' \le T_n$ and $j \le i_n - 1$ unless S queries $x \in \{0,1\}^{\le T_n^{1/2}}$ such that $F_x \in V_{|x|,i_n-1}$ to $\mathcal{V}$. Note that, if $n' \in \mathbb{N}$ satisfies $n' < t^{-1}(T_n^{1/2})$, then we have $i_{\max}(n') < \log\log(T_n^{1/2}) = \log\log T_n - 1 = i_n - 1$. Thus, $V_{n',i_n-1} = \emptyset$. Otherwise, $V_{|x|,i_n-1}$ is selected from $V_{|x|,i_n-2}$ uniformly at random. Thus, such a conditional probability is bounded above by

$$\begin{aligned} T_n^{1/2} \max_{t^{-1}(T_n^{1/2}) \le n' \le T_n^{1/2}} \frac{\lfloor p(n')|V_{n',i_n-2}|\rfloor}{|V_{n',i_n-2}|} &= O\left(T_n^{1/2} p(t^{-1}(T_n^{1/2}))\right) \\ &= O\left(T_n^{1/2} \cdot (T_n^{1/2})^{-\sqrt{\epsilon(n)}}\right) \\ &= T_n^{-\omega(1)} \\ &= n^{-\omega(1)}, \end{aligned}$$

where the last equation holds because $T_n \geq n$.

Under the condition that the given instance $x \in \{0,1\}^n$ is determined by only $\rho_{n',j}$ for $n' \leq T_n$ and $j \leq i_n - 1$, the depth $d+2$ circuit C constructed during the execution of $\mathcal{A}(\langle M, d\rangle, x, 1^{T_n^2})$ is determined only by $\rho_{n',j}$ for $n' \leq T_n$ and $j \leq i_n - 1$. Then, applying the restriction $\rho_{n',j}$ for $n' \leq T_n$ and $j \leq i_n$ under this condition is regarded as a $p(n')$-random restriction for V_{n',i_n-1} for each $n' \leq T_n$, where we can ignore small n' such that $n' < t^{-1}(T_n)$ because $i_{\max}(n') < \log\log T_n = i_n$ for such n'. Note that the width and internal size of C are at most T_n and 2^{T_n+1}, respectively. Thus, by applying Lemma 8, the probability that C does not become a constant (i.e., the probability that $\mathcal{A}$ returns "?") is at most

$$\begin{aligned} O\left(T_n^2 \max_{t^{-1}(T_n)\leq n'\leq T_n} p(n')^{1/(d+2)}\ell(n')^2\right) &= O\left(T_n^2 \max_{t^{-1}(T_n)\leq n'\leq T_n} t(n')^{-\frac{\omega(1)}{d+2}+4}\right) \\ &= T_n^{-\omega(1)} \\ &= n^{-\omega(1)}, \end{aligned}$$

where the last equation holds because $T_n \geq n$. □

6.4 Oracle Separation between UP ∩ coUP and Distributional Σ_d^{p}

Theorem 12. *For any constant $a > 0$ and $d \in \mathbb{N}$, there exists an oracle $\mathcal{O}_{a,d}$ satisfying (1) $\mathsf{Dist}\Sigma_d^{p\mathcal{O}_{a,d}} \subseteq \mathsf{AvgP}^{\mathcal{O}_{a,d}}$ and (2) $\mathsf{UP}^{\mathcal{O}_{a,d}} \cap \mathsf{coUP}^{\mathcal{O}_{a,d}} \not\subseteq \mathsf{BPTIME}^{\mathcal{O}_{a,d}}[2^{\frac{an}{\log n}}]$.*

Proof sketch. Let $c = \max\{21(d+2)a, 1\}$ and $\epsilon(n) = 1/4a$ for each $n \in \mathbb{N}$. We construct an oracle $\mathcal{O}_{a,d}$, as in Section 6.1, where we set the parameters as follows:

$$t(n) = 2^{\frac{n}{\epsilon(n)\cdot\log n}},\ p(n) = t(n)^{-5(d+2)},\ \ell(n) = t(n)^2,\ \text{and}\ i_{\max}(n) = \frac{1}{c}\log\log t(n).$$

We also change the simulation overhead from T^2 to T^{2^c} and the setting of i from $\log\log n$ to $c^{-1}\log\log n$ in $\mathcal{A}$. Then, we can easily show the analog of Lemma 9. Further, we get

$$p(n)^{i_{\max}(n)-1} \geq t(n)^{-\frac{5(d+2)\log n}{c}} = 2^{-\frac{20(d+2)an\log n}{c\log n}} \geq 2^{-\frac{20}{21}n},$$

and $|V_{n,i_{\max}(n)}| = \Omega(p(n)^{i_{\max}(n)-1}\cdot 2^n) = 2^{\Omega(n)} > 0$. Thus, we can show the hardness of computing in $t(n) = 2^{\frac{n}{\epsilon(n)\cdot\log n}} = 2^{4an/\log n}$ steps using the same proof as that of Theorem 10. It is not hard to verify that this lower bound yields $\mathsf{UP}^{\mathcal{O}_{a,d}} \cap \mathsf{coUP}^{\mathcal{O}_{a,d}} \not\subseteq \mathsf{BPTIME}^{\mathcal{O}_{a,d}}[2^{\frac{an}{\log n}}]$ as Corollary 2. The average-case easiness of $\mathsf{Dist}\Sigma_d^p$ also holds by the same argument to as proof of Theorem 11, where we select T_n as the maximum value of n^4, n^{2^c}, the 2^c-th power of the time for $S^?$ to generate an instance of size n, and the 2^c-th power of the time to execute $M^?$ on instance size n (see also Section 7.3). □

7 Oracle Separation: Learning and Distributional PH

We prove the oracle separation between the hardness of learning and distributional PH. In Sections 7.1– 7.3, we present Theorem 13. Note that we can obtain the second item of Theorem 2 as a corollary to Theorem 13 (i.e., Corollary 4). The first item of Theorem 2 is shown in a similar way by changing the parameters (we will discuss this in Section 7.4).

Theorem 13. *Let $a : \mathbb{N} \to \mathbb{R}_{>0}$ be a function such that $n/\omega(\log^2 n) \le a(n) \le O(1)$. For any $d \in \mathbb{N}$ and sufficiently large $c \in \mathbb{N}$, there exists an oracle $\mathcal{O} := \mathcal{O}_{a,c,d}$ such that (1) $\mathsf{Dist}\Sigma_d^{p\mathcal{O}} \subseteq \mathsf{AvgP}^{\mathcal{O}}$ and (2) $\mathsf{SIZE}^{\mathcal{O}}[n]$ is not weakly PAC learnable with membership queries in time $2^{\frac{a(n)n}{\log n}}$ on $\mathcal{D}$, where $\mathcal{D}$ is an arbitrary class of example distributions such that $\mathcal{D}_n$ contains all uniform distributions over subsets $S \subseteq \{0,1\}^n$ with $|S| \ge 2^{(1-\frac{a(n)}{c})\cdot n}$.*

Corollary 4. *For any constants $a, \epsilon > 0$ and $d \in \mathbb{N}$, there exists an oracle $\mathcal{O}$ such that (1) $\mathsf{Dist}\Sigma_d^{p\mathcal{O}} \subseteq \mathsf{AvgP}^{\mathcal{O}}$ and (2) $\mathsf{SIZE}^{\mathcal{O}}[n]$ is not weakly PAC learnable with membership queries in $2^{\frac{an}{\log n}}$ steps on all uniform distributions over subsets $S \subseteq \{0,1\}^n$ with $|S| \ge 2^{(1-\epsilon)\cdot n}$.*

7.1 Construction of Random Oracle

Let $\epsilon\colon \mathbb{N} \to \mathbb{N}$ and $c, d \in \mathbb{N}$ denote parameters such that $\Omega(1) \le \epsilon(n) \le n/\omega(\log^2 n)$ and $c \ge \max\{3, 26(d+2)/\epsilon(n)\}$ for sufficiently large n.

Construction. $\mathcal{O}_{\epsilon,c,d} = \mathcal{F} + \mathcal{A}$, where each oracle is randomly selected by the following procedure:

1. Define functions t, p, ℓ, q, and $i_{\max}$ by

$$t(n) = 2^{\frac{n}{\epsilon(n)\cdot\log n}},\ p(n) = t(n)^{-11(d+2)},\ \ell(n) = t(n)^4, q(n) = t(n)^{-3(d+2)},\ \text{and } i_{\max}(n) = \frac{1}{c}\log\log t(n).$$

2. For each $n \in \mathbb{N}$, define a set $V_{n,0}$ of variables on alphabet Σ_n of size $\ell(n)$ by

$$V_{n,0} = \{F_z : z \in \{0,1\}^n\}.$$

   We assume that each alphabet in Σ_n has a binary representation of length at most $\lceil \log \ell(n) \rceil$.

3. For each $n \in \mathbb{N}$, define a set $W_{n,0}$ of variables on alphabet $\{0,1\}$ by

$$W_{n,0} = \{G_{z,x} : z, x \in \{0,1\}^n\}.$$

4. For each $n \in \mathbb{N}$ and $i \in [i_{\max}(n) - 1]$, we inductively (on i) define a $p(n)$-random restriction $\rho_{n,i}$ to $V_{n,i-1}$ and a $q(n)$-random restriction $\sigma_{n,i}$ to $W_{n,i-1}$, and we define subsets $V_{n,i} \subset V_{n,i-1}$ and $W_{n,i} \subset W_{n,i-1}$ of variables by

$$V_{n,i} = \{v \in V_{n,i-1} : \rho_{n,i}(v) = *\} \text{ and } W_{n,i} = \{w \in W_{n,i-1} : \sigma_{n,i}(w) = *\}.$$

   We also define $\rho_{n,i_{\max}(n)}$ (resp. $\sigma_{n,i_{\max}(n)}$) as a 0-random restriction (i.e., a full assignment) to $V_{n,i_{\max}(n)-1}$ (resp. $W_{n,i_{\max}(n)-1}$). Let $\rho_{n,i} \equiv \rho_{n,i_{\max}(n)}$ and $\sigma_{n,i} \equiv \sigma_{n,i_{\max}(n)}$ for $i \ge i_{\max}(n)+1$. For simplicity, we may identify $\rho_{n,i}$ (resp. $\sigma_{n,i}$) with a composite restriction $\rho_{n,1}\cdots\rho_{n,i}$ to $V_{n,0}$ (resp. $\sigma_{n,1}\cdots\sigma_{n,i}$ to $W_{n,0}$) for each n and i.

5. Let $f = \{f_n\}_{n\in\mathbb{N}}$, where $f_n\colon \{0,1\}^n \to \Sigma_n$ is a random function defined by $f_n(z) = \rho_{n,i_{\max}(n)}(F_z)$. Let $g = \{g_n\}_{n\in\mathbb{N}}$, where $g_n\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ is a random function defined by $g_n(z,x) = \sigma_{n,i_{\max}(n)}(G_{z,x})$.

6. Define $\mathcal{F} = \{\mathcal{F}_n\}_{n\in\mathbb{N}}$, where $\mathcal{F}_n\colon \{0,1\}^n \times \Sigma_n \times \{0,1\}^n$ as follows:

$$\mathcal{F}_n(z,y,x) = \begin{cases} g_n(z,x) & \text{if } y = f_n(z) \\ 0 & \text{otherwise.} \end{cases}$$

7. Define $\mathcal{A}$ as follows: On input $(\langle M, d\rangle, x, 1^{T^{2^c}})$, where M is an oracle machine, $d \in \mathbb{N}$, $x \in \{0,1\}^*$, and $T \in \mathbb{N}$, the oracle $\mathcal{A}$ returns the value in $\{0,1,?\}$ determined according to the following procedure:

   1: let $i := \frac{1}{c}\log\log T$;
   2: construct a depth $d+2$ circuit C corresponding to the quantified formula

$$\exists w_1 \in \{0,1\}^{|x|} \forall w_2 \in \{0,1\}^{|x|}, \ldots, Q_d w_d \in \{0,1\}^{|x|}, M^{\mathcal{O}}(x, w_1, w_2, \ldots, w_d),$$

   where $Q_d = \exists$ if d is an odd number; otherwise, $Q_d = \forall$.
   First, we construct a depth-d circuit that represents the above-mentioned quantified formula whose leaf corresponds to $M^{\mathcal{O}}(x, w_1, w_2, \ldots, w_d)$ for some $w_1, w_2, \ldots, w_d$, where we truncate $w_1, w_2, \ldots, w_d$ into a string of length T because we will execute M in only T steps. Then, we replace each leaf with a DNF formula of width $2T$ to obtain the circuit C, where each term corresponds to one possible choice of $\mathcal{F}$ such that $M^{\mathcal{F}+\mathcal{A}}(x, w_1, w_2, \ldots, w_d)$ halts with an accepting state after execution in T steps. In other words, we arbitrarily consider functions $f' = \{f'_n\}_{n=1}^T$ and $g' = \{g'_n\}_{n=1}^T$, where $f'_n \colon \{0,1\}^n \to \Sigma_n$ and $g'_n \colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$, define an oracle $\mathcal{F}'$ in the same manner as $\mathcal{F}$, and execute $M^{\mathcal{F}'+\mathcal{A}}(x, w_1, w_2, \ldots, w_d)$ in T steps. If M queries (z, y, x) to $\mathcal{F}'$, then we add literals to the corresponding term in the following manner:

$$\text{add literals} \begin{cases} (F_z \neq y) & \text{if } f'(z) \neq y \\ (F_z = y) \text{ and } (G_{z,x} = 0) & \text{if } f'(z) = y \text{ and } g'(z,x) = 0 \\ (F_z = y) \text{ and } (G_{z,x} = 1) & \text{if } f'(z) = y \text{ and } g'(z,x) = 1 \end{cases}$$

   By the construction, the above-mentioned quantified formula is satisfied when $M^{\mathcal{F}+\mathcal{A}}$ is executed in T steps iff C returns 1 when it is restricted by $\rho_{1,i_{\max}(1)}, \ldots, \rho_{T,i_{\max}(T)}$ and $\sigma_{1,i_{\max}(1)}, \ldots, \sigma_{T,i_{\max}(T)}$. We can also easily verify that the width of C is at most $2T$ and the internal size of C is at most 2^{T+1}.
   3: if $C|_{\rho_{1,i},\ldots,\rho_{T,i}} \equiv b$ for some $b \in \{0,1\}$, then **return** b; otherwise, **return** "?".

To verify that $\mathcal{A}$ is not circular on recursive calls for $\mathcal{A}$, it is sufficient to check the following:

Lemma 11. *For each input, the value of $\mathcal{A}(\langle M, d\rangle, x, 1^{T^{2^c}})$ is determined by only $\rho_{n,j}$ and $\sigma_{n,j}$ for $n \leq T$ and $j \leq \frac{1}{c}\log\log T$ $(= i)$.*

Proof. We show the lemma by induction on T. We consider the execution of $(\langle M, d\rangle, x, 1^{T^{2^c}})$. We remark that $\mathcal{A}$ first makes a depth $d+2$ circuit C based on M, and C is independent of the value of $\mathcal{F}$.

We assume that M makes some valid query $(\langle M', d'\rangle, x', 1^{T'^{2^c}})$ to $\mathcal{A}$ recursively on constructing C. Since the length of such a query is at most T, we have $T'^{2^c} \leq T$. If we let $i' = c^{-1}\log\log T'$, then we have

$$i' = \frac{1}{c}\log\log T' \leq \frac{1}{c}\log\log T^{\frac{1}{2^c}} = \frac{1}{c}\log\log T - 1.$$

By the induction hypothesis, the recursive answer of $\mathcal{A}$ is determined by only $\rho_{n,j}$ and $\sigma_{n,j}$ for $n \leq T'$ and $j \leq i-1$, and so is C. The lemma holds because the answer of $\mathcal{A}$ is determined by restricting C by $\rho_{n,j}$ and $\sigma_{n,j}$ for $n \leq T$ and $j \leq i$. □

Note that we may omit the subscripts ϵ, c, and d from $\mathcal{O}_{\epsilon,c,d}$.

7.2 Hardness of Learning

Theorem 14. *For arbitrary parameters $\epsilon(n)$, c, and d such that $\Omega(1) \leq \epsilon(n) \leq n/\omega(\log^2 n)$ and $c \geq \max\{3, 26(d+2)/\epsilon(n)\}$ (for sufficiently large n), the following event occurs with probability 1 over the choices of $\mathcal{O} := \mathcal{O}_{\epsilon,c,d}$: a concept class $\mathscr{C}^{\mathcal{O}}$ defined by*

$$\mathscr{C}^{\mathcal{O}} = \{\mathcal{F}_n(z, y, \cdot) : n \in \mathbb{N}, z \in \{0,1\}^n, y \in \Sigma_n\}$$

is not weakly PAC learnable (with confidence error at most $1/3$) in $t(n) = 2^{\frac{n}{\epsilon(n)\log n}}$ steps on $\mathcal{D}$, where $\mathcal{D}$ is a class of example distributions such that $\mathcal{D}_n$ contains all uniform distributions over subsets $S \subseteq \{0,1\}^n$ with $|S| \geq 2^{(1-\frac{4(d+2)}{c\epsilon(n)})\cdot n}$.

In the remainder of Section 7.2, we present the formal proof of Theorem 14.

For any choice of $\mathcal{O}$ and $z \in \{0,1\}^n$, we define a subset $G^*_z \subseteq \{0,1\}^n$ by

$$G^*_z = \{x \in \{0,1\}^n : \rho_{n,i_{\max}(n)-1}(G_{z,x}) = *\}.$$

We let U^*_z denote a uniform distribution over the elements in G^*_z. For $z \in \{0,1\}^n$, we define a function $g_z : \{0,1\}^n \to \{0,1\}$ by $g_z(x) = g(z,x)$ $(= \mathcal{F}_n(z, f_n(z), x) \in \mathscr{C}^{\mathcal{O}})$. We introduce the notion of hard indices as follows.

Definition 7. *We say that $z \in \{0,1\}^n$ is a hard index if $z \in V_{n,i_{\max}(n)-1}$ and $|G^*_z| \geq 2^{(1-\frac{4(d+2)}{c\epsilon(n)})\cdot n}$.*

We show that f_z is hard to learn on example distribution U^*_z for a hard index z.

First, we estimate the probability that such a hard index exists.

Lemma 12. *If $\epsilon(n) \geq \Omega(1)$ and $c \geq 26(d+2)/\epsilon(n)$, then we have*

$$\Pr_{\mathcal{O}}[\textit{there exists no hard index in } \{0,1\}^n] \leq n^{-\omega(1)}.$$

Proof. Since $\epsilon(n) \geq \Omega(1)$, we have $t(n) \leq 2^n$ and $i_{\max}(n) \leq \frac{1}{c}\log n$ for sufficiently large n. Thus, we have that for sufficiently large n,

$$p(n)^{i_{\max}(n)} \geq t(n)^{-\frac{11(d+2)}{c}\log n} = 2^{-\frac{11(d+2)n\log n}{c\epsilon(n)\log n}} \geq 2^{-\frac{11}{26}n},$$

and

$$q(n)^{i_{\max}(n)} \geq t(n)^{-\frac{3}{c}\log n} = 2^{-\frac{3n\log n}{c\epsilon(n)\log n}} \geq 2^{-\frac{3}{26}n}.$$

By Lemma 3, for any $z \in \{0,1\}^n$, we get

$$\begin{aligned}
\Pr\left[\sum_{z\in V_{n,i_{\max}(n)-1}} |G^*_z| < \frac{1}{2}\cdot\frac{|V_{n,i_{\max}(n)-1}|\cdot 2^n}{2^{2n}}|W_{n,i_{\max}(n)-1}|\right] &< 2\exp\left(-\Omega(p(n)^{2i_{\max}(n)}q(n)^{i_{\max}(n)})\cdot 2^{2n}\right)\\
&\leq 2\exp\left(-\Omega(2^{-\frac{22}{26}n}\cdot 2^{-\frac{3}{26}n}\cdot 2^{2n})\right)\\
&\leq \exp(-2^{\Omega(n)})\\
&= n^{-\omega(1)}.
\end{aligned}$$

If the above-mentioned event does not occur, then there exists $z \in V_{n,i_{\max}(n)-1}$ such that

$$|G^*_z| \geq \frac{|W_{n,i_{\max}(n)-1}|}{2^{n+1}} = \Omega(q(n)^{i_{\max}(n)}\cdot 2^n) = \Omega(2^{n-\frac{3(d+2)n}{c\epsilon(n)}}).$$

Thus, $|G^*_z| \geq 2^{n-\frac{4(d+2)n}{c\epsilon(n)}}$ for sufficiently large $n \in \mathbb{N}$, and such z is a hard index. □

Now, we fix a (randomized) learning algorithm L and $n \in \mathbb{N}$ arbitrarily. For Theorem 14, by the Borel–Cantelli lemma and the countability of uniform learners, it is sufficient to show the following: for sufficiently large $n \in \mathbb{N}$ and $\delta_n = t(n)^{-1/4}$ ($\geq n^{-\omega(1)}$),

$$\Pr_{\mathcal{O}}[\text{for all } z \in \{0,1\}^n \text{ and } D \in \mathcal{D}_n,$$
$$\Pr_{L,S}\left[L^{\mathcal{O},g_z}(S) \to h^{\mathcal{O}} \text{ s.t. } \Pr_{x \leftarrow D}[h^{\mathcal{O}}(x) = g_z(x)] \geq \frac{1}{2} + \delta_n \text{ within } t(n) \text{ steps } \right] \geq \frac{2}{3}\Bigg] \leq n^{-\omega(1)}, \quad (4)$$

where S is a sample set of size at most $t(n)$ generated according to $EX(g_z, D)$.

For $z \in \{0,1\}^n$, $x \in \{0,1\}^n$, and a sample set S, we use the notation $L^{\mathcal{O},g_z}(S)(x)$ to refer to the following procedure: (1) execute $L^{\mathcal{O},g_z}(S)$; (2) if L outputs some hypothesis h within $t(n)$ steps, then execute $h^{\mathcal{O}}(x)$. For $z \in \{0,1\}^n$, we define events I_z and J_z (over the choice of $\mathcal{O}$) as follows:

$$I_z = \left(\Pr_{L,S,x}\left[\mathcal{F}(z, f_n(z), x') \text{ is queried for some } x' \in \{0,1\}^n \text{ during } L^{\mathcal{O},g_z}(S)(x)\right] \geq \delta_n^4\right)$$
$$J_z = \left(\Pr_{L,S}\left[L^{\mathcal{O},g_z}(S) \to h^{\mathcal{O}} \text{ s.t. } \Pr_{x}[h^{\mathcal{O}}(x) = g_z(x)] \geq \frac{1}{2} + \delta_n \text{ within } t(n) \text{ steps } \right] \geq \frac{2}{3}\right),$$

where S is selected according to $EX(g_z, U_z^*)$, and x is selected according to U_z^*.

We assume that $z \in \{0,1\}^n$ is a hard index. Then, we have $|G_z^*| \geq 2^{(1-\frac{4(d+2)}{c\epsilon(n)})\cdot n}$; thus, U_z^* must be contained in $\mathcal{D}_n$. Therefore, the left-hand side of the inequality (4) is bounded above by

$$\begin{aligned}
\Pr_{\mathcal{O}}\left[\bigwedge_{z:\text{hard}} J_z\right] &\leq \Pr_{\mathcal{O}}\left[\bigwedge_{z:\text{hard}} J_z \vee I_z\right] \\
&\leq \Pr_{\mathcal{O}}\left[\left(\bigwedge_{z:\text{hard}} I_z\right) \vee \left(\bigvee_{z:\text{hard}} J_z \wedge \neg I_z\right)\right] \\
&\leq \Pr_{\mathcal{O}}\left[\bigwedge_{z:\text{hard}} I_z\right] + \Pr_{\mathcal{O}}\left[\bigvee_{z:\text{hard}} J_z \wedge \neg I_z\right]. \quad (5)
\end{aligned}$$

Here, we let P_1 and P_2 represent the first and second terms of (5), respectively. We derive the upper bounds on P_1 and P_2 as the following lemmas, which immediately imply the inequality (4).

Lemma 13. $P_1 = \Pr_{\mathcal{O}}\left[\bigwedge_{z:hard} I_z\right] \leq n^{-\omega(1)}$.

Lemma 14. $P_2 = \Pr_{\mathcal{O}}\left[\bigvee_{z:hard} J_z \wedge \neg I_z\right] \leq n^{-\omega(1)}$.

7.2.1 Proof of Lemma 13

First, we fix random restrictions except for $\rho_{n,i_{\max}(n)}$ and use π to denote the composite restriction. Note that all the hard indices are determined at this stage. Assume that there exists a hard index of length n, and let $z_\pi \in \{0,1\}^n$ be the lexicographically first hard index. Then, we can divide $\rho_{n,i_{\max}(n)}$ into two random selections as follows. First, we randomly select unassigned values of $f_n(z)$ except for $f_n(z_\pi)$ (let π' denote the corresponding random restriction). Then, we select the remaining value of $f_n(z_\pi)$ from Σ_n uniformly at random.

We remark that π determines g and G_z^* for all $z \in \{0,1\}^n$. Now, we construct a randomized oracle machine A to compute $f_n(z_\pi)$ based on L, g, G_z^*, π, π', and additional oracle access to $\mathcal{V}$, where $\mathcal{V}(y)$ returns 1 if $y = f_n(z_\pi)$ (otherwise, returns 0).

On input z_π and oracle access to $\mathcal{V}$, A executes L in $t(n)$ steps for a target function g_z and an example distribution U_z^* (note that examples and membership queries are simulated by g and G_z^*); if L outputs some hypothesis h, then compute $h(x)$ for $x \leftarrow U_z^*$, where A answers the queries of L and h to $\mathcal{O}$ as follows:

> $\mathcal{F}(z, y, x)$: If $z = z_\pi$, then A queries y to $\mathcal{V}$; if $\mathcal{V}$ returns 1, then return $g_n(z, x)$ (otherwise, return 0). In other cases, A can correctly answer $\mathcal{F}(z, y, x)$ because it is determined by π and π'.
> $\mathcal{A}(\langle M, d\rangle, x, 1^{T^{2^c}})$: Since A executes L only $t(n)$ steps, we can assume that the size of h is at most $t(n)$ and it is evaluated in time $O(t(n)^2)$. Thus, we can assume that $T^{2^c} = O(t(n)^2)$ and for sufficiently large n,
>
> $$i = \frac{1}{c}\log\log T = \frac{1}{c}\log\log O(t(n))^{1/2^{c-1}} \leq \frac{1}{c}\log\log t(n)^{1/2^{c-2}} \leq \frac{1}{c}\log\log t(n) - \frac{c-2}{c},$$
>
> which is strictly smaller than $i_{\max}(n)$ $(= \frac{1}{c}\log\log t(n))$ because $c \geq 3$. By Lemma 11, the answer does not depend on $\rho_{n,i_{\max}(n)}$. Thus, A can correctly simulate $\mathcal{A}$ by π and π' in this case.

A repeats the above-mentioned executions of L and its hypothesis n/δ_n^4 times. If A queries y such that $\mathcal{V}(y) = 1$ at some trial, then A outputs y $(= f_n(z_\pi))$ and halts (otherwise, A outputs $\perp$).

By the construction, A can correctly simulate L and its hypothesis h for a target function g_z and an example distribution U_z^*. It is easy to verify that the number q of the queries of A to $\mathcal{V}$ is bounded as $q \leq (n/\delta_n^4) \cdot O(t(n)^2) \leq O(n) \cdot t(n)^3$.

Assume that $\wedge_{z:\mathrm{hard}} I_z$ holds. Then, L or h queries $(z_\pi, f_n(z_\pi), \cdot)$ to $\mathcal{F}$ with a probability of at least δ_n^4 for each trial. Since A repeats this trial n/δ_n^4-times, the failure probability of A is at most $(1-\delta_n^4)^{n/\delta_n^4} < 2^{-n}$. Thus, we have

$$\Pr_{\mathcal{O},A}\left[A^{\mathcal{V}}(z_\pi) = f_n(z_\pi) \middle| \pi, \pi', \wedge_{z:\mathrm{hard}} I_z\right] \geq 1 - 2^n. \tag{6}$$

Meanwhile, even under the condition on π and π', the value of $f_n(z_\pi)$ is selected from Σ_n at random independently of A. Thus, we can also show that

$$\Pr_{\mathcal{O},A}\left[A^{\mathcal{V}}(z_\pi) = f_n(z_\pi) \middle| \pi, \pi', \wedge_{z:\mathrm{hard}} I_z\right] \leq \frac{q}{\ell(n)} = \frac{O(n)}{t(n)} = n^{-\omega(1)}. \tag{7}$$

The above-mentioned inequality (7) contradicts the inequality (6). This indicates that there exists no hard index in this case. By Lemma 12, we conclude that

$$P_1 = \Pr_{\mathcal{O}}\left[\bigwedge_{z:\mathrm{hard}} I_z\right] \leq \Pr_{\mathcal{O}}\left[\text{there exists no hard index in } \{0,1\}^n\right] \leq n^{-\omega(1)}.$$

7.2.2 Proof of Lemma 14

We fix $z \in \{0,1\}^n$ arbitrarily, and we let $\mathcal{O}'$ denote a partial choice of $\mathcal{O}$ except for the values of $g(z, x)$, where $x \in G_z^*$. Then, we have

$$\Pr_{\mathcal{O}}\left[\neg I_z \wedge J_z\right] = \mathrm{E}_{\mathcal{O}'}\left[\Pr_{\mathcal{O}}\left[\neg I_z \wedge J_z | \mathcal{O}'\right]\right].$$

We remark that g_z is a truly random (partial) function on G_z^*, even under the condition on $\mathcal{O}'$.

Assume that z is a hard index, and $\neg I_z \wedge J_z$ occurs. Let $N = |G_z^*|$. Since z is a hard index, $N \geq 2^{(1-\frac{4(d+2)}{c\epsilon(n)})\cdot n} \geq 2^{\Omega(n)}$ holds.

By Markov's inequality and $\neg I_z$, we have

$$\Pr_{L,S}\left[\Pr_x\left[\mathcal{F} \text{ is asked } (z, f_n(z), \cdot) \text{ during } L^{\mathcal{O},g_z}(S)(x)\right] \leq 4\delta_n^3\right] \geq 1 - \frac{\delta_n}{4}.$$

By J_z, we also get

$$\Pr_{L,S}\left[L^{\mathcal{O},g_z}(S) \to h^{\mathcal{O}} \text{ s.t. } \Pr_x[h^{\mathcal{O}}(x) = g_z(x)] \geq \frac{1}{2} + \delta_n \text{ within } t(n) \text{ steps }\right] \geq \frac{2}{3}.$$

By the two above-mentioned inequalities, there exist a sample set S and a random string r for L such that

- $L^{\mathcal{O},g_z}(S;r)$ outputs some hypothesis $h^{\mathcal{O}}$ in time $t(n)$ without querying $(z, f_n(z), \cdot)$ to $\mathcal{F}$;
- $\Pr_{x \leftarrow U_z^*}[h^{\mathcal{O}}(x) \text{ queries } (z, f_n(z), \cdot) \text{ to } \mathcal{F}] \leq 4\delta_n^3$; and
- $\Pr_{x \leftarrow U_z^*}[h^{\mathcal{O}}(x) = f_z(x)] \geq \frac{1}{2} + \delta_n$.

If L and h do not query $(z, f_n(z), \cdot)$ to $\mathcal{F}$ and they halt in $t(n)$ steps, then the answers by $\mathcal{O}$ do not depend on $\sigma_{n,i_{\max}(n)}$, i.e., the values of $g_z(x)$ for $x \in G_z^*$, as seen in the proof of Lemma 13. In other words, they are totally determined by $\mathcal{O}'$. Thus, we can replace $\mathcal{O}$ with $\mathcal{O}'$ in these cases (where we assume that $\mathcal{O}'$ returns an error on an undefined input).

Now, we show that a truth table $\tau \in \{0,1\}^N$ of g_z on G_z^* has a short description (under the condition on $\mathcal{O}'$), which yields the upper bound on P_2 because a random function does not have such a short description with high probability.

Let $B_z \subseteq G_z^*$ be the subset consisting of x such that $h^{\mathcal{O}}(x)$ queries $(z, f_n(z), \cdot)$ to $\mathcal{F}$. By the second property, we have $|B_z| \leq 4N\delta_n^3$. We consider the following reconstruction procedure for τ. First, we execute $L^{\mathcal{O}',g_z}(S;r)$ to obtain $h^{\mathcal{O}'}$. Note that if we obtain all answers for membership queries by L as auxiliary advice Q (of length at most $t(n)$), then we can remove external access to g_z from L. Next, we execute $h^{\mathcal{O}'}(x)$ on each input $x \in G_z^* \setminus B_z$. By combining these predictions with auxiliary advice $S_B = \{(x, g_z(x)) : x \in B_z\}$, we also obtain a partial truth table $\tilde{\tau} \in \{0,1\}^N$ $(1/2 - \delta_n)$-close to $\tau \in \{0,1\}^N$. If we obtain $err \in \{0,1\}^N$ defined by $err_i = \tau_i \oplus \tilde{\tau}_i$ as auxiliary advice, then we can reconstruct τ from $\tilde{\tau}$ and err.

Therefore, we can reconstruct τ from L, S, r, Q, S_B, and err under the condition on $\mathcal{O}'$. Since the Hamming weight of err is at most $N \cdot (1/2 + \delta_n)$, err is represented by a binary string of length at most $(1 - \Omega(\delta_n^2)) \cdot N$ by lexicographic indexing among binary strings of the same weight. Hence, τ has a short description of length at most

$$O(t(n)) + 4\delta_n^3(n+1) \cdot N + (1 - \Omega(\delta_n^2)) \cdot N \leq O(t(n)) + \left(1 - \Omega(\delta_n^2)\right) \cdot N.$$

Since τ is a truly random string even under the condition on $\mathcal{O}'$, we have

$$\begin{aligned}
\Pr_{\mathcal{O}}\left[z \text{ is hard and } \neg I_z \wedge J_z | \mathcal{O}'\right] &\leq \frac{2^{O(t(n)) + \left(1-\Omega(\delta_n^2)\right)\cdot N}}{2^N} \\
&\leq 2^{O(t(n)) - \Omega(t(n)^{-1/2})\cdot N} \\
&\leq 2^{2^{O(n/\log n)} - 2^{\Omega(n - n/\log n)}} \\
&\leq 2^{-2^{\Omega(n)}}.
\end{aligned}$$

This implies that $\Pr_{\mathcal{O}}[z \text{ is hard and } \neg I_z \wedge J_z] \leq \mathrm{E}_{\mathcal{O}'}[\Pr_{\mathcal{O}}[z \text{ is hard and } \neg I_z \wedge J_z | \mathcal{O}']] \leq 2^{-2^{\Omega(n)}}$ for any index z. Note that the number of indices is at most 2^n. Thus, by taking the union bound, we conclude that

$$P_2 = \Pr_{\mathcal{O}}\left[\bigvee_{z:\text{hard}} J_z \wedge \neg I_z\right] \leq 2^n \cdot 2^{-2^{\Omega(n)}} = n^{-\omega(1)}.$$

7.3 Average-Case Easiness of $\Sigma_{\mathrm{d}}^{\mathrm{p}}$

Theorem 15. *For any parameters $\epsilon(n)$, c, and d such that $\Omega(1) \leq \epsilon(n) \leq n/\omega(\log^2 n)$ and $c \geq \max\{3, 26(d+2)/\epsilon(n)\}$ (for sufficiently large n), the following event occurs with probability 1 over the choice of $\mathcal{O} := \mathcal{O}_{\epsilon,c,d}$: for all tuples of a polynomial-time oracle machine $M^?$ and a polynomial-time randomized oracle sampling machine $S^?$, there exists a deterministic polynomial-time errorless heuristic oracle machine with a failure probability of at most n^{-2} for the distributional Σ_d^p problem $(L_M^{\mathcal{O}}, D_S^{\mathcal{O}})$, defined as follows: $(D_S^{\mathcal{O}})_n \equiv S^{\mathcal{O}}(1^n)$ for each $n \in \mathbb{N}$ and*

$$L_M^{\mathcal{O}} = \{x \in \{0,1\}^* : \exists w_1 \in \{0,1\}^{|x|} \forall w_2 \in \{0,1\}^{|x|}, \ldots, Q_d w_d \in \{0,1\}^{|x|}, M^{\mathcal{O}}(x, w_1, w_2, \ldots, w_d)\},$$

where $Q_d = \exists$ if d is an odd number; otherwise, $Q_d = \forall$.

By a simple padding argument on the instance size and the argument in [Imp95, Proposition 3], we obtain the following corollary to Theorem 15.

Corollary 5. *Let $\epsilon(n)$, c, and d denote the same parameters as in Theorem 15. With probability 1 over the choice of $\mathcal{O} := \mathcal{O}_{\epsilon,c,d}$, the event $\mathsf{Dist}\Sigma_d^{p\mathcal{O}} \subseteq \mathsf{AvgP}^{\mathcal{O}}$ occurs.*

Theorem 13 immediately follows from Theorem 14 and Corollary 5 by selecting $\epsilon(n) = 1/a(n)$ and sufficiently large c for ϵ^{-1} and d.

Proof of Theorem 15. For each n, let T_n be the maximum value of n^{2^c}, the 2^c-th power of the time for $S^?$ to generate an instance of size n, and the 2^c-th power of the time to execute $M^?$ on input of size n. Let $i_n = c^{-1} \log\log T_n$.

Now, we construct an errorless heuristic scheme $B^{\mathcal{O}}$ that is given $x \in \{0,1\}^n$ as input and returns a value of $\mathcal{A}(\langle M, d\rangle, x, 1^{T_n^{2^c}})$. Note that $B^{\mathcal{O}}(x) = L_M^{\mathcal{O}}(x)$ unless $\mathcal{A}(\langle M, d\rangle, x, 1^{T_n^{2^c}})$ outputs "?". Thus, we will show the inequality

$$P_{n,M,S} := \Pr_{\mathcal{O},S}\left[\mathcal{A}(\langle M, d\rangle, x, 1^{T_n^{2^c}}) = \text{"?" where } x \leftarrow S^{\mathcal{O}}(1^n)\right] \leq O\left(\frac{1}{n^4}\right). \tag{8}$$

Then, by applying Markov's inequality, we have

$$\Pr_{\mathcal{O}}\left[\Pr_S\left[B^{\mathcal{O}}(x) = L_M^{\mathcal{O}}(x) \text{ where } x \leftarrow S^{\mathcal{O}}(1^n)\right] \geq \frac{1}{n^2}\right] \leq O\left(\frac{1}{n^2}\right),$$

and the theorem follows from the Borel–Cantelli lemma and the countability of (M, S).

To show the inequality (8), we first show that the instance $x \in \{0,1\}^n$ is determined by only $\rho_{n',j}$ for $n' \leq T_n$ and $j \leq i_n - 1$ with a probability of at least $1 - O(n^{-4})$. Then, we will show that $\mathcal{A}(M, x, 1^{T_n^{2^c}})$ returns $M^{\mathcal{O}}(x)$ with a probability of at least $1 - O(n^{-4})$ under the condition that x is determined by only $\rho_{n',j}$ for $n' \leq T_n$ and $j \leq i_n - 1$.

By the same argument as the proof of Theorem 11, the first probability is bounded above by

$$\begin{aligned} T_n^{1/2^c} O\left(\max_{t^{-1}(T_n^{1/2^c})\le n'\le T_n^{1/2^c}} \{p(n'), q(n')\}\right) &= O\left(T_n^{1/2^c} q(t^{-1}(T_n^{1/2^c}))\right) \\ &= O\left(T_n^{1/2^c} \cdot (T_n^{1/2^c})^{-3(d+2)}\right) \\ &= O((T_n^{1/2^c})^{-(3d+5)}) \\ &= O(n^{-4}), \end{aligned}$$

where the last equation holds because $T_n \ge n^{2^c}$.

By the same argument as the proof of Theorem 11, the second probability that $\mathcal{A}$ returns "?" under the condition that the given instance $x \in \{0,1\}^n$ is determined by only $\rho_{n',j}$ for $n' \le T_n$ and $j \le i_n - 1$ is at most

$$\begin{aligned} O\left(T_n^2 \max_{t^{-1}(T_n)\le n'\le T_n} \{p(n')^{1/(d+2)}\ell(n')^2, q(n')^{1/(d+2)} \cdot 2^2\}\right) &= O\left(T_n^2 \max_{t^{-1}(T_n)\le n'\le T_n} t(n')^{-3}\right) \\ &= O(T_n^{-1}) \\ &= O(n^{-4}), \end{aligned}$$

where the last equation holds because $T_n \ge n^{2^c} \ge n^4$. □

7.4 Oracle Separation between Learning and Distributional PH

Theorem 16. *For any function ϵ such that $\omega(1) \le \epsilon(n) \le n/\omega(\log^2 n)$ and an arbitrary small constant $\delta \in (0,1)$, there exists an oracle $\mathcal{O}$ such that (1) $\mathsf{DistPH}^{\mathcal{O}} \subseteq \mathsf{AvgP}^{\mathcal{O}}$ and (2) $\mathsf{SIZE}^{\mathcal{O}}[n]$ is not weakly PAC learnable with membership queries in time $2^{O(\frac{n}{\epsilon(n)\log n})}$ on $\mathcal{D}$, where $\mathcal{D}$ is an arbitrary class of example distributions such that $\mathcal{D}_n$ contains all uniform distributions over subsets $S \subseteq \{0,1\}^n$ with $|S| \ge 2^{(1-\epsilon(n)^{-(1-\delta)})\cdot n}$.*

Proof sketch. Let $c = 3$. We construct an oracle $\mathcal{O}$, as in Section 7.1, where we set the parameters as follows:

$$t(n) = 2^{\frac{n}{\epsilon(n)\cdot\log n}},\ p(n) = t(n)^{-\epsilon(n)^\delta},\ \ell(n) = t(n)^2, q(n) = t(n)^{-\epsilon(n)^\delta},\ \text{and}\ i_{\max}(n) = c^{-1}\log\log t(n).$$

Then, the hardness of learning follows by the same argument as the proof of Theorem 14, where we can select the lower bound of G_z^* for a hard index z as

$$|G_z^*| \ge 2^{n-1} \cdot q(n)^{i_{\max}(n)} \ge 2^{(1-\frac{1}{3\epsilon(n)^{1-\delta}})n+1} \ge 2^{(1-\epsilon(n)^{-(1-\delta)})n},$$

for sufficiently large n. The average-case easiness of DistPH also holds by essentially the same argument as the proofs of Theorems 11 and 15. □

Acknowledgment

The authors would like to thank the anonymous reviewers for many helpful comments. Shuichi Hirahara was supported by JST, PRESTO Grant Number JPMJPR2024, Japan. Mikito Nanashima was supported by JST, ACT-X Grant Number JPMJAX190M, Japan.

References

[ABX08] B. Applebaum, B. Barak, and D. Xiao. On Basing Lower-Bounds for Learning on Worst-Case Assumptions. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, FOCS'08, pages 211–220, 2008.

[AF09] Luis Filipe Coelho Antunes and Lance Fortnow. Worst-Case Running Times for Average-Case Algorithms. In *Proceedings of the Conference on Computational Complexity (CCC)*, pages 298–303, 2009.

[AFvV06] L. Antunes, L. Fortnow, D. van Melkebeek, and N. Vinodchandran. Computational depth: Concept and applications. *Theoretical Computer Science*, 354(3):391–404, 2006. Foundations of Computation Theory (FCT 2003).

[AGGM06] A. Akavia, O. Goldreich, S. Goldwasser, and D. Moshkovitz. On Basing One-Way Functions on NP-Hardness. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, STOC ' 06, pages 701 – –710, New York, NY, USA, 2006. ACM.

[AGvM+18] E. Allender, J. A. Grochow, D. van Melkebeek, C. Moore, and A. Morgan. Minimum Circuit Size, Graph Isomorphism, and Related Problems. *SIAM Journal on Computing*, 47(4):1339–1372, 2018.

[BB15] A. Bogdanov and C. Brzuska. On Basing Size-Verifiable One-Way Functions on NP-Hardness. In *Theory of Cryptography - 12th Theory of Cryptography Conference, TCC 2015, Warsaw, Poland, March 23-25, 2015, Proceedings, Part I*, pages 1–6, 2015.

[BCGL92] Shai Ben-David, Benny Chor, Oded Goldreich, and Michael Luby. On the Theory of Average Case Complexity. *J. Comput. Syst. Sci.*, 44(2):193–219, 1992.

[BEHW87] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Occam's Razor. *Inf. Process. Lett.*, 24(6):377–380, apr 1987.

[BFP05] Harry Buhrman, Lance Fortnow, and Aduri Pavan. Some Results on Derandomization. *Theory Comput. Syst.*, 38(2):211–227, 2005.

[BL13] A. Bogdanov and C. Lee. Limits of Provable Security for Homomorphic Encryption. In *Advances in Cryptology - CRYPTO 2013 - 33rd Annual Cryptology Conference, Santa Barbara, CA, USA, August 18-22, 2013. Proceedings, Part I*, pages 111–128, 2013.

[BT06a] A. Bogdanov and L. Trevisan. Average-Case Complexity. *Foundations and Trends in Theoretical Computer Science*, 2(1):1 – –106, 2006.

[BT06b] A. Bogdanov and L. Trevisan. On Worst-Case to Average-Case Reductions for NP Problems. *SIAM J. Comput.*, 36(4):1119 – –1159, December 2006.

[CIKK16] M. Carmosino, R. Impagliazzo, V. Kabanets, and A. Kolokolova. Learning Algorithms from Natural Proofs. In *Proceedings of the 31st Conference on Computational Complexity*, CCC'16. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2016.

[CIKK17] M. Carmosino, R. Impagliazzo, V. Kabanets, and A. Kolokolova. Agnostic Learning from Tolerant Natural Proofs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*, volume 81 of *LIPIcs*, pages 35:1–35:19, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Dan16] A. Daniely. Complexity Theoretic Limitations on Learning Halfspaces. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC'16, pages 105–117, New York, NY, USA, 2016. ACM.

[DP12] I. Damgård and S. Park. Is Public-Key Encryption Based on LPN Practical? *IACR Cryptology ePrint Archive*, 2012:699, 2012.

[DSS16] A. Daniely and S. Shalev-Shwartz. Complexity Theoretic Limitations on Learning DNF's. In *Proceedings of 29th Conference on Learning Theory*, volume 49 of *COLT'16*, pages 815–830, Columbia University, New York, USA, 23–26 Jun 2016. PMLR.

[FF93] J. Feigenbaum and L. Fortnow. Random-Self-Reducibility of Complete Sets. *SIAM Journal on Computing*, 22(5):994–1005, 1993.

[GV08] D. Gutfreund and S. Vadhan. Limitations of Hardness vs. Randomness under Uniform Reductions. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques. APPROX 2008, RANDOM 2008*, volume 5171 of *LNCS*, pages 469–482, 2008.

[Hir18] S. Hirahara. Non-Black-Box Worst-Case to Average-Case Reductions within NP. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 247–258, 2018.

[Hir20] S. Hirahara. Non-Disjoint Promise Problems from Meta-Computational View of Pseudorandom Generator Constructions. In *35th Computational Complexity Conference (CCC 2020)*, volume 169 of *LIPIcs*, pages 20:1–20:47, Dagstuhl, Germany, 2020.

[Hir21] S. Hirahara. Average-Case Hardness of NP from Exponential Worst-Case Hardness Assumptions. In *53rd Annual ACM Symposium on Theory of Computing (STOC 2021)*, 2021.

[HMX10] I. Haitner, M. Mahmoody, and D. Xiao. A New Sampling Protocol and Applications to Basing Cryptographic Primitives on the Hardness of NP. In *IEEE 25th Annual Conference on Computational Complexity*, pages 76–87, 2010.

[HS17] Shuichi Hirahara and Rahul Santhanam. On the Average-Case Complexity of MCSP and Its Variants. In *Proceedings of the Computational Complexity Conference (CCC)*, pages 7:1–7:20, 2017.

[HW20] S. Hirahara and O. Watanabe. On Nonadaptive Security Reductions of Hitting Set Generators. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2020, August 17-19, 2020, Virtual Conference*, volume 176 of *LIPIcs*, pages 15:1–15:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[IL89] R. Impagliazzo and M. Luby. One-way Functions Are Essential for Complexity Based Cryptography. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, pages 230–235, 1989.

[IL90] R. Impagliazzo and L. Levin. No better ways to generate hard NP instances than picking uniformly at random. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, FOCS'90, pages 812–821, 1990.

[ILO20] R. Ilango, B. Loff, and I. C. Oliveira. NP-Hardness of Circuit Minimization for Multi-Output Functions. In *Proceedings of the 35th Computational Complexity Conference*, CCC '20. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2020.

[Imp95] R. Impagliazzo. A personal view of average-case complexity. In *Proceedings of IEEE Tenth Annual Conference on Structure in Complexity Theory*, pages 134–147, 1995.

[Imp11] R. Impagliazzo. Relativized Separations of Worst-Case and Average-Case Complexities for NP. In *2011 IEEE 26th Annual Conference on Computational Complexity*, pages 104–114, 2011.

[IW97] Russell Impagliazzo and Avi Wigderson. $P = BPP$ if E Requires Exponential Circuits: Derandomizing the XOR Lemma. In *Proceedings of the Symposium on the Theory of Computing (STOC)*, pages 220–229, 1997.

[KL18] P. Kothari and R. Livni. Agnostic Learning by Refuting. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 55:1–55:10, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[KSS94] M. Kearns, R. Schapire, and L. Sellie. Toward Efficient Agnostic Learning. *Machine Learning*, 17(2):115–141, Nov 1994.

[LV91] M. Li and P. Vitányi. Learning Simple Concepts under Simple Distributions. *SIAM Journal on Computing*, 20(5):911–935, 1991.

[LV16] T. Liu and V. Vaikuntanathan. On Basing Private Information Retrieval on NP-Hardness. In *Theory of Cryptography - 13th International Conference, TCC 2016-A, Tel Aviv, Israel, January 10-13, 2016, Proceedings, Part I*, pages 372–386, 2016.

[LV19] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, Cham, 2019.

[MU05] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, USA, 2005.

[Nan21] M. Nanashima. A Theory of Heuristic Learnability. In *Proceedings of the 34th Conference on Learning Theory, COLT'21*. PMLR, 2021.

[PV88] L. Pitt and L. Valiant. Computational Limitations on Learning from Examples. *J. ACM*, 35(4):965–984, October 1988.

[Raz93] A. Razborov. An Equivalence between Second Order Bounded Domain Bounded Arithmetic and First Order Bounded Arithmetic. In *Arithmetic, Proof Theory and Computational Complexity*, 1993.

[Reg09] O. Regev. On Lattices, Learning with Errors, Random Linear Codes, and Cryptography. *J. ACM*, 56(6), September 2009.

[RR97] Alexander A. Razborov and Steven Rudich. Natural Proofs. *J. Comput. Syst. Sci.*, 55(1):24–35, 1997.

[Sch90] R. Schapire. The Strength of Weak Learnability. *Mach. Learn.*, 5(2):197–227, 1990.

[Vad17] S. Vadhan. On learning vs. refutation. In *Proceedings of the 2017 Conference on Learning Theory (COLT'17)*, volume 65 of *Proceedings of Machine Learning Research*, pages 1835–1848, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

[Val84] L. Valiant. A Theory of the Learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[Wat12] Thomas Watson. Relativized Worlds without Worst-Case to Average-Case Reductions for NP. *ACM Trans. Comput. Theory*, 4(3), September 2012.

[Xia09] D. Xiao. On basing ZK $\neq$ BPP on the hardness of PAC learning. In *Proceedings of the 24th Conference on Computational Complexity*, CCC'09, pages 304–315, 2009.

A Relativized Worst-Case-to-Average-Case Connections for PH

In this appendix, we observe that some of the results from [Hir21] can be relativized.

Theorem 17. *For every oracle A, and for every constant $k \geq 1$, if $\mathsf{Dist}\Sigma_{k+1}^{\mathrm{p}^A} \subseteq \mathsf{AvgP}^A$, then $\Sigma_k^{\mathrm{p}^A} \subseteq \mathsf{DTIME}(2^{O(n/\log n)})^A$.*

[Hir21] presented a non-relativizing proof of Theorem 17. The only non-relativizing part of the proof of [Hir21] is the pseudorandom generator construction of Buhrman, Fortnow, and Pavan [BFP05] under the assumption that NP is easy on average. Their proof uses a variant of the PCP theorem, which is known to be non-relativizing.

Theorem 18 (Buhrman, Fortnow, and Pavan [BFP05])**.** *If $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, then $\mathsf{E} \not\subseteq \mathsf{i.o.SIZE}(2^{\epsilon n})$ for some constant $\epsilon > 0$.*

Here, we present a relativizing proof of a weaker statement.

Theorem 19. *For every oracle A, if $\mathsf{DistP}^{\mathsf{NP}^A} \subseteq \mathsf{AvgP}^A$, then $\mathsf{E}^A \not\subseteq \mathsf{i.o.SIZE}(2^{\epsilon n})^A$ for some constant $\epsilon > 0$.*

Proof. The proof consists of two parts. First, we show that there exists a P^A-natural property useful against $\mathsf{SIZE}(2^{n/2})^A$. Second, using the natural property, we present a strongly exponential lower bound for $\mathsf{E}^{\mathsf{NP}^A}$. This is sufficient because $\mathsf{E}^{\mathsf{NP}^A} = \mathsf{E}^A$ holds under the assumption that $\mathsf{P}^{\mathsf{NP}^A}$ is easy on average, which was proven in [BCGL92].

For a string $x \in \{0,1\}^*$, let $\mathrm{size}^A(x)$ denote the minimum size of an A-oracle circuit whose truth table is equal to x. Following [HS17], we claim that there is a P^A-natural property useful against $\mathsf{SIZE}(2^{n/2})^A$. Let $L := \{x \in \{0,1\}^* \mid \mathrm{size}^A(x) \leq |x|^{0.5}\}$. (This is the Minimum Circuit Size Problem with size parameter $2^{n/2}$.) It is easy to observe that $L \in \mathsf{NP}^A$. Consider the uniform distribution $\mathcal{U} = \{\mathcal{U}_n\}_{n\in\mathbb{N}}$. Since $(L,\mathcal{U}) \in \mathsf{DistNP}^A \subseteq \mathsf{AvgP}^A$, there exists an errorless heuristic polynomial-time algorithm M^A such that $\Pr_{x\in\{0,1\}^n}[M^A(x) = L(x)] \geq \frac{3}{4}$ and $M^A(x) \in \{1, \perp\}$ for every $x \in \{0,1\}^*$ such that $\mathrm{size}^A(x) < |x|^{1/2}$. The number of Yes instances in L is small: by a standard counting argument, it can be shown that $\Pr_{x\in\{0,1\}^n}[L(x) = 1] = o(1)$. Thus, M^A outputs 0 for at least a $\frac{3}{4} - o(1)$ fraction of the inputs, whereas it outputs either 1 or $\perp$ on any Yes instance of L.

Next, we claim that $\mathsf{E}^{\mathsf{NP}^A} \not\subseteq \mathsf{i.o.SIZE}(2^{n/2})$. By a standard search-to-decision algorithm for NP, there exists a $\mathsf{P}^{\mathsf{NP}^A}$ algorithm H^A that, on input $N \in \mathbb{N}$ represented in unary, finds the lexicographically first string $f \in \{0,1\}^N$ such that $M^A(f) = 0$. Note that there exists such a string f. Moreover, $M^A(f) = 0$ implies that $\mathrm{size}^A(f) \geq |f|^{1/2}$. Now, consider the following $\mathsf{E}^{\mathsf{NP}^A}$ algorithm: On input $x \in \{0,1\}^*$ of length $n \in \mathbb{N}$, simulate H^A on input 2^n to obtain the truth table

$f \in \{0,1\}^{2^n}$ and output the x-th bit of f. Since this algorithm computes a function whose truth table is f, we conclude that $\mathsf{E}^A = \mathsf{E}^{\mathsf{NP}^A} \not\subseteq \mathsf{i.o.SIZE}(2^{n/2})^A$. □

A consequence of Theorem 19 is that there exists a pseudorandom generator secure against linear-sized circuits.

Corollary 6. *For every oracle A, if $\mathsf{DistP}^{\mathsf{NP}^A} \subseteq \mathsf{AvgP}^A$, then there exists a pseudorandom generator*

$$G = \{G_n : \{0,1\}^{O(\log n)} \to \{0,1\}^n\}$$

secure against A-oracle linear-sized circuits and computable in time $n^{O(1)}$ with oracle access to A. In particular, $\mathsf{P}^A = \mathsf{BPP}^A$.

Proof. This follows from Theorem 19 and the fact that the theorem of Impagliazzo and Wigderson [IW97] relativizes. □

Now, we sketch the proof of Theorem 17. For simplicity, we consider only the following special case.

Theorem 20. *For every oracle A, if $\mathsf{DistΣ}_2^{\mathrm{p}A} \subseteq \mathsf{AvgP}^A$, then $\mathsf{NP}^A \subseteq \mathsf{DTIME}(2^{O(n/\log n)})^A$.*

Proof Sketch. The proof consists of three steps.

1. If $\mathsf{DistΣ}_2^{\mathrm{p}A} \subseteq \mathsf{AvgP}^A$, then $\mathsf{Gap}(\mathrm{K}^{\mathsf{NP}^A}\text{ vs }\mathrm{K}^A) \in \mathsf{P}^A$.
2. If $\mathsf{Gap}(\mathrm{K}^{\mathsf{NP}^A}\text{ vs }\mathrm{K}^A) \in \mathsf{P}^A$, then every language in NP^A admits an A-oracle universal heuristic scheme.
3. Any language with an A-oracle universal heuristic scheme can be solved in time $2^{O(n/\log n)}$ with oracle access to A.

The first two steps use the existence of a pseudorandom generator, which follows from Corollary 6. (In the original proof of [Hir21], the non-relativizing proof of Theorem 18 was used.) It is not hard to observe that the proofs of the three steps listed above are relativizing. □

B Proof of Lemma 3

We can identify a random choice of n elements from U with n consecutive random choices of one element from U, where the chosen element is removed from U. We remark that these n choices are dependent on the previous choices, and we cannot directly apply the Chernoff bound. Instead, we apply the martingale theory. The basic background can be founded elsewhere [MU05].

For each $i \in \mathbb{N}$, let X_i be a random variable that returns 1 if the i-th chosen element is contained in S and 0 otherwise. Let $m = \sum_{i=1}^{n} X_i$. Then, the statement in the lemma is written as follows:

$$\Pr_{X_1,\ldots,X_n}\left[\left|m - \frac{M}{N}n\right| > \gamma \cdot \frac{M}{N}n\right] < 2e^{-2\gamma^2\cdot(\frac{M}{N})^2\cdot n}.$$

For each $i \in \mathbb{N} \cup \{0\}$, we define Z_i by

$$Z_i = \frac{M - \sum_{k=1}^{i} X_k}{N - i}n.$$

First, we show that these $Z_0, Z_1, \ldots, Z_n$ constitute a martingale.

Claim 1. *The sequence of $Z_0, Z_1, \ldots, Z_n$ is a martingale with respect to $X_1, \ldots, X_n$.*

Proof. It is sufficient to show that for each i, $\mathrm{E}[Z_{i+1}|X_1, \ldots, X_i] = Z_i$.

Fix $X_1, \ldots, X_{i-1}$ arbitrarily, where $i \leq n$. Let $X = \sum_{k=1}^{i} X_k$. If $X = M$, then $\mathrm{E}[Z_{i+1}|X_1, \ldots, X_i] = 0 = Z_i$. Even when $X < M$, the same equation holds as follows:

$$\begin{aligned}
\mathrm{E}[Z_{i+1}|X_1, \ldots, X_i] &= n \cdot \left[\frac{M - X - 1}{N - i - 1} \cdot \Pr[X_{i+1} = 1|X] + \frac{M - X}{N - i - 1} \cdot \Pr[X_{i+1} = 0|X] \right] \\
&= n \cdot \left[\frac{M - X - 1}{N - i - 1} \cdot \frac{M - X}{N - i} + \frac{M - X}{N - i - 1} \cdot \frac{(N - M) - (i - X)}{N - i} \right] \\
&= n \cdot \frac{(M - X)(N - i - 1)}{(N - i - 1)(N - i)} \\
&= n \cdot \frac{M - X}{N - i} \\
&= Z_i.
\end{aligned}$$

□

Thus, the sequence of $Z_0, Z_1, \ldots, Z_n$ is a martingale (with respect to themselves).

For each $i \leq n$, under the condition on $X_1, \ldots, X_{i-1}$, we have

$$\frac{M - \sum_{k=1}^{i-1} X_k - 1}{N - i} n \leq Z_i \leq \frac{M - \sum_{k=1}^{i-1} X_k}{N - i} n.$$

Thus, we can show that

$$\begin{aligned}
Z_i - Z_{i-1} &\leq \frac{M - \sum_{k=1}^{i-1} X_k}{N - i} n - \frac{M - \sum_{k=1}^{i-1} X_k}{N - i + 1} n \\
&= \frac{M - \sum_{k=1}^{i-1} X_k}{(N - i)(N - i + 1)} n,
\end{aligned}$$

and

$$\begin{aligned}
Z_i - Z_{i-1} &\geq \frac{M - \sum_{k=1}^{i-1} X_k - 1}{N - i} n - \frac{M - \sum_{k=1}^{i-1} X_k}{N - i + 1} n \\
&= \frac{M - \sum_{k=1}^{i-1} X_k - (N - i + 1)}{(N - i)(N - i + 1)} n
\end{aligned}$$

Therefore, if we define a new random variable B_i by

$$B_i = \frac{M - \sum_{k=1}^{i-1} X_k - (N - i + 1)}{(N - i)(N - i + 1)} n,$$

then we get

$$B_i \leq Z_i - Z_{i-1} \leq B_i + \frac{n}{N - i}.$$

Now, we apply the Azuma-Hoeffding inequality for $Z_0, \ldots, Z_n$. Then, for any real value $\lambda \geq 0$, we have

$$\begin{aligned}
\Pr\left[|Z_n - Z_0| > \lambda\right] &< 2 \exp\left(- \frac{2\lambda^2}{\sum_{i=1}^{n} (\frac{n}{N-i})^2} \right) \\
&\leq 2 \exp\left(- \frac{2\lambda^2 (N - n)^2}{n^3} \right).
\end{aligned}$$

Note that $Z_n = \frac{M-\sum_{i=1}^{n} X_i}{N-n} n = \frac{M-m}{N-n} n$ and $Z_0 = \frac{M}{N} n$. If we assume that $|m - \frac{M}{N} n| > \gamma \frac{M}{N} n$, then it is not hard to verify that

$$|Z_n - Z_0| > \gamma \cdot \frac{Mn^2}{N(N-n)}.$$

Therefore, by applying the above-mentioned inequality for $\lambda = \gamma \cdot \frac{Mn^2}{N(N-n)}$, we conclude that

$$\begin{aligned}
\Pr\left[\left|m - \frac{M}{N} n\right| > \gamma \cdot \frac{M}{N} n\right] &\leq \Pr\left[|Z_n - Z_0| > \gamma \cdot \frac{Mn^2}{N(N-n)}\right] \\
&< 2 \exp\left(-\frac{2(\frac{\gamma M n^2}{N(N-n)})^2 (N-n)^2}{n^3}\right) \\
&= 2 \exp\left(-2\gamma^2 \cdot (\frac{M}{N})^2 \cdot n\right).
\end{aligned}$$

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il